INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UNIVERSITÉ D'OTTAWA UNIVERSITY OF OTTAWA

ÉCOLE DES ÉTUDES SUPÉRIEURES ET DE LA RECHERCHE

SCHOOL OF GRADUATE STUDIES AND RESEARCH

KORIR, Daniel
AUTEUR DE LA THÈSE-*AUTHOR OF THESIS*

M.A. (Education)
GRADE-*DEGREE*

EDUCATION
FACULTÉ, ÉCOLE, DÉPARTEMENT-*FACULTY, SCHOOL, DEPARTMENT*

TITRE DE LA THÈSE-*TITLE OF THE THESIS*

THE EFFECTS OF ITEM DIFFICULTY AND EXAMINEE ABILITY ON THE DISTRIBUTION AND EFFECTIVENESS OF LZ AND ECIZ4 APPROPRIATENESS INDICES

M. Boss
DIRECTEUR DE LA THÈSE-*THESIS SUPERVISOR*

EXAMINATEURS DE LA THÈSE-*THESIS EXAMINERS*

D. Laveault

D. Zimmerman

(LE DOYEN DE L'ÉCOLE DES ÉTUDES SUPÉRIEURES ET DE LA RECHERCHE) SIGNATURE (DEAN OF THE SCHOOL OF GRADUATE STUDIES AND RESEARCH)

THE EFFECTS OF ITEM DIFFICULTY AND EXAMINEE ABILITY ON THE DISTRIBUTION AND EFFECTIVENESS OF LZ AND ECIZ4 APPROPRIATENESS INDICES.

Daniel K. Korir

Thesis submitted to the School of Graduate studies and Research in partial fulfilment of the requirements for the Masters Degree in Education.

University of Ottawa

ABSTRACT

Test scores are intended to provide a measure of examinee's estimate of ability. High ability examinees are expected to get few easy items wrong and low ability examinees are exepcted to get few difficult items right. But there are occasions when the test-taking behavior of some atypical examinees may be so unsual that their test scores cannot be regarded as an appropriate measure of ability. An atypical examinee can have a spuriously low or a spuriously high score.

However, appropriateness indices can be used to identify examinees with potentially inaccurate total scores. Appropriateness indices provide quantitative measures of response pattern atypicality. These indices fall into two major categories: (a) IRT-based and (b) non-IRT based indices. The dependency of non-IRT based indices on the item difficulty order of a particular group has rendered them inadequate for detecting aberrant reponse patterns. IRT-based indices are group invariant.

Researchers have investigated the effectiveness and the distributions of these indices under varying conditions of testing. However, some test situations might require efficient and accurate indices of appropriateness measurement for restricted samples. It might be helpful, for example, to accurately identify examinees with potential spuriously low scores falling just the below the criterion of a minimum competency test. On a certification test, it might be helpful to concentrate on identifying examinees with

spuriously high scores. Therefore, the effects of item difficulty and examinee ability distributions on the effectiveness and the distributional characteristics of LZ and ECIZ4 (IRT-based) appropriateness indices were investigated in this study.

To examine the effects of item difficulty and ability distributions on the distributional characteristics of LZ and ECIZ4, data were generated in nine combinations of item difficulty and ability distributions to simulate the responses of 2000 examinees to 60 test items according to the three-parameter model. Three uniform distributions of item difficulty were used. Items typical of diagnostic tests were generated in the interval -3.0 to +1.2; items typical of power tests were generated in the interval of -3.0 to +3.0; and items typical of certification and licencing tests were generated in the interval of -1.2 to +3.0. Three distributions of ability were used. Thetas typical of low, medium, and high ability examinees were generated to have normal distributions with the means of -1.2, 0.0, and +1.2 respectively and each with a standard deviation of 0.6.

The mean, standard deviation, skewness, kurtosis, and the percentile estimates of LZ and ECIZ4 were significantly affected by the variations of item difficulty and ability distributions. The distributions of the two indices approximated a normal distribution when the ability estimates matched the item difficulty. Overall, the distributions of LZ approximated a normal distribution better than the distribution of ECIZ4.

To examine the effectiveness of LZ and ECIZ4 in detecting

aberrant response patterns, two samples, each consisting of 500 response patterns (for spuriously low and spuriously high) were generated for each of the nine combinations of item difficulty and ability distribution and subjected to spurious treatments. Twenty percent and 10% spuriously high scores were created by randomly selecting 20% or 10% of the original responses and changing incorrect answers to correct. Twenty percent and 10% spuriously low scores were created by randomly selecting 20% or 10% of the original responses and changing correct answers to incorrect. The percentile estimates obtained were used as cutoff points to classify response patterns as aberrant or non-aberrant.

Spuriously low aberrant response patterns were easier to detect by the two indices under the low item difficulty and spuriously high aberrant response patterns were easier to detect under high item difficulty. At low (0.01 and 0.05) false positive rates, LZ had higher detection rates of spuriously high and spuriously low aberrant response patterns than ECIZ4 under the high item difficulty; and ECIZ4 had higher detection rates than LZ under the medium and under the low item difficulty. Twenty percent treatment samples were easier to detect by the two indices than the 10% treatment samples.

ACKNOWLEDGEMENTS.

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

TABLE OF CONTENTS.

LIST OF TABLES.

CHAPTER I

INTRODUCTION

A test is a systematic procedure for measuring a sample of examinee behavior. In the strictest sense, a test measures only test-taking behavior, that is, the responses a person makes to the test items. A person is not measured directly; rather a person's characteristics (traits) are inferred from his or her responses to a test. If the behaviors exhibited on the test adequately mirror the construct being measured, the test will provide useful information. If the test does not adequately reflect the underlying characteristics, inferences made from test scores are inappropriate.

Test scores can only be useful in estimating person ability if the person's pattern of responses to the items corresponds to his or her expected response pattern. For instance, if the test consists of k dichotomous items arranged in ascending order of difficulty (from easy to difficult), then someone who gets x of the k items right is expected to have answered the first x items correctly and the last k-x items incorrectly. If it is the easy items that he or she gets wrong, his or her pattern is regarded as deviating from the expected pattern. Therefore, a score with such a response pattern is said to be inappropriate in estimating the person's ability.

There are many factors that can make a person's response

pattern inappropriate. Among them is how clearly the instructions are understood by the examinee, familiarity with test materials and with the concepts used, previous experience with test tasks or with similar tasks and with working under pressure, and motivational factors (Van der Flier, 1982). Birenbaum (1985) notes different causes of aberrant (unexpected) response patterns, misconceptions concerning the subject matter, cultural bias, test anxiety, exceptional creativity, lack of concentration resulting in carelesly reading the questions, guessing, and occasional copying a more able neighbour's work. Wright (1977) mentions tendencies such as sleeping, fumbling, and plodding as causes of unusual response patterns. He defines sleeping as those examinees who get bored with a test and do poorly in the beginning because of confusion with test format. Examinees who never get to the latter items on the test are plodders. Unusual response patterns can also result from technical problems such as answer sheet alignment.

However, these factors jeopardise the validity of the response patterns and they are not directly reflected by a total test score. Checking the validity of the response pattern, therefore, becomes a necessity for ensuring an accurate assessment of performance. This validity check of response patterns is done with the help of appropriateness indices which provide automated means for identifying response patterns where total test score may provide misleading information.

Several indices for detecting aberrant (unusual) response patterns have been developed. These indices describe the degree to

which an indivitual's pattern of item responses is unusual. These indices can be classified into two groups. One group consists of indices based solely on the actual observed response patterns of the group of examinees. Examples of these indices include Sato's caution index (1975), Van der flier's U''' index (1982), Donlon and Fischer's personal biserial (1968), Tatsuoka and Tatsuoka's norm conformity index (1982), and Harnish and Linn's modified caution index (1981). The other group consists of indices based solely on IRT models. Examples of these indices are the fit indices developed by Wright and his associates (1977), the appropriateness indices developed by Levine and Rubin (1979), and the group of extended caution indices developed by Tatsuoka and Linn (1983).

Most of the previous researchers in appropriateness measurement have compared the effectiveness of appropriateness indices (Levine & Rubin, 1979; Rudner, 1983; Parsons, 1983; Birenbaum, 1985; Drasgow, Levine, & Williams, 1985, 1986, 1987; Tomsic, 1986; Noonan, 1990; Candell, 1990); others have investigated the distribution of appropriateness indices under different conditions of item and ability parameters (Molenaar & Hoijtink, 1990; Hoijtink, 1986; and Drasgow, 1985). In the next section, five group dependent indices are presented. The performance of these indices under varying conditions of testing has been investigated by previous researchers . Further, these indices have also been found to be useful in practical testing situations (Harnisch & Linn, 1981; Jaeger, 1988; Kellett, 1989).

Non-IRT Indices

Sato's caution index is an example of a non-IRT index. Computations involving this index entail the construction of an s-p (student problem) chart which consists of a binary data matrix in which the students (represented by rows) are arranged from top to bottom in descending order of their total test scores, and test items are arranged from left to right in ascending order of difficulty. The caution index, C, for student i is defined as:

$$C = 1 - \frac{COV\ (X_{ij}\ ,\ n_j)}{COV\ (U_{ij}\ ,\ n_j)}$$

where X_{ij} is student i's response (1 or 0) to item j; U_{ij} is item-j score for a hypothetical ''ideal'' student with the same total score of X_i as student i; n_j is the number of students correctly answering item j; and COV(,) denotes the covariance (over items) between the two variables inside the parentheses. This index ranges from 0 (for the ideal student) to appproximately 1 (or more than 1) for a student whose response pattern consists of random a sequence of 1's and 0's (Sato, 1975).

Van der Flier's U''' index (Van der Flier, 1982) uses an item deviance function of 0-1 (0 for wrong, 1 for right) to calculate examinees deviant scores. The model assumes that the scores of the items are stochastically independent and the item characteristic curves increase monotonically. The item characteristic curve shows

the proportion of respondents who answer the item correctly as a function of their position on the latent (ability) variable. To determine the probabilities of different score patterns, the item parameters are estimated from a reference group. The deviant index U''' is defined as:

$$U'''(X) = \frac{\log p \max - \log p(x)}{\log p \max - \log p \min}$$

where p(x) is the probability of pattern X, and p max and p min indicate the probabilities of the most and the least deviant pattern yielding the same score. The items are arranged in ascending order of difficulty (from easy to difficult).

Like Van der Flier's U''' index, Donlon and Fischer's personal biserial (r) requires that items be arranged in ascending order of difficulty (Donlon & Fischer, 1968). Personal biserial is defined as the biserial correlation between a person's distribution of item responses on a specific test and the distribution of item difficulties generated by some reference group.

The personal biserial index essentially measures the relationship between the difficulty of the items in the test for the person as evidenced by his/her passes and failures and the difficulty of the items as evidenced by the group determined item difficulties. It is positive when there is agreement. A person who tends to get difficult items correct (through guessing or creative

thinking) and easy items incorrect, thus disagreeing with the whole group, would have a small or even a negative r. Similarly, scores which resulted from pure chance responses would generate personal biserials with a mean of zero, because responses made in this manner would not be correlated with difficulty.

The Norm Conformity Index (NCI) measures the extent to which a given individual's response pattern resembles a reversed Guttman vector (where 1's precede all 0's) with the same number of 1's. Items are arranged in ascending order of difficulty for an arbitrary norm group or the group for which the individual is a member (Tatsuoka & Tatsuoka, 1982). The NCI is defined as:

$$NCI_i = 2U_{ia} / U_i - 1$$

Where U_{ia} is the sum of the number of 0's to the right of each 1 in the vector Y_i(response vector for examinee i) added over all 1's and U_i is the product of the number of 1's and the number of 0's which represents the number of (0,1) and (1,0) pairs in vector Y_i. The lower limit of -1, is attained when the individual's response pattern forms a Guttman vector, with all 0's preceding all 1's while an upper limit of 1.0 is attained when the response pattern of the individual forms a ''reversed Guttman vector'', with the ones coming first (Tatsuoka & Tatsuoka, 1982).

The modified caution index, MCI, based on Sato's (1975) index was proposed by Harnisch and Linn (1981). It is obtained from the matrix of an examinee's item responses (0 for wrong, 1 for right)

by group-determined item difficulties (p-values) arranged in ascending order. Like the previous indices, an individual is expected to get the easiest items correct and hard ones incorrect. The MCI shows the extent to which the individual's score pattern deviates from the Guttman scale. The larger the index, the greater the departure from the Guttman scale and the less likely the pattern reflects ''normal'' performance across items. The modified caution index has an upper limit of one and lower limit of zero. The MCI for the i-th examinee is given by :

$$MCI = \frac{\sum_{J=1}^{N_{i.}} (1 - U_{ij}) N_{.j} - \sum_{J=N_{i.}+1}^{J} U_{ij} N_{.j}}{\sum_{J=1}^{N_{i.}} N_{.j} - \sum_{J=J+I-n_{.j}}^{J} N_{.j}}$$

where i indexes the examinee; j indexes the item; U_{ij} is the response of examinee i to item j (0 for wrong, 1 for right); $N_{.j}$ is the total number of correct responses to the j^{th} item and $N_{i.}$ is the total correct for the i^{th} examinee. By being bound between zero and one, the MCI eliminates extreme scores that are sometimes obtained with Sato's caution index, especially in cases where a very high scoring examinee misses a single very easy item.

In summary non-IRT indices rely solely on observed item

responses and statistics such as item difficulty. They are very useful measures in specific situations such as with teacher-made tests or with the assessment of curriculum or instructional effects. For instance, Tatsuoka and Birenbaum (1979) used the Norm Conformity Index to demonstrate that student's errors on certain types of arithmetic problems are frequently systematic. Jaeger (1988) used MCI in a standard setting procedure to identify judges with aberrant judgment. Kellett (1989) used the MCI along with other indices to assess the measurement equivalence of different Canadian sub-population (cultural) groups. The MCI and the NCI have been found to be the most useful indices among non-IRT appropriatenes indices because they are not influenced by extreme scores. They are inexpensive to compute too.

IRT-Based Indices

The appropriateness indices in this section are based on item response theory (IRT). Item response theory comprises a class of models that assess the relationship between an individual's responses to items on a test or a questionnaire and the psychological trait measured by the instrument. In other words, IRT models the relation between an individual's standing on a latent characteristic (attitude, ability, etc) and the item responses.

A number of IRT-based appropriateness indices have been proposed and they can be sub-divided into (1) unstandardized and standardized extended caution indices, (2) maximum likelihood indices, and (3) person fit indices.

Extended Caution Indices. Extended caution indices have been developed from Sato's caution index. In the extended caution indices, the ideal response curves are replaced by examinee response curves theoretically derived from IRT. The response curve for examinee i is obtained by holding θ as a constant and considering b (item difficulty) as a continuous variable in the logistic function. Intuitively, the examinee response curve at a fixed level of θ corresponds to a step function whose values equals one for $b > \theta$ and zero for $b \leq \theta$. The six extended caution indices which have been developed are ECI1, ECI2, ECI3, ECI4, ECI5, and ECI6.

The second and the fourth extended caution indices have been found to be very useful. They are defined as:

$$ECI2 = 1 - \frac{COV\ (\ Y_i\ ,\ G_j\ )}{COV\ (\ P_i\ ,\ G_j\ )}$$

$$\text{and } ECI4 = 1 - \frac{COV\ (\ Y_i\ ,\ P_i\ )}{COV\ (\ G_j\ ,\ P_i\ )}$$

where Y_{ij} is the observed binary matrix; P_{ij} is the probability of subject i correctly answering item j according to the one-, two-, or three-parameter model; Y_i is the vector of binary scores for subject i or the i^{th} row vector; P_i is the probability vector from

the i^{th} row in the probability matrix, and G_j (referring to item j) is the j^{th} element of a vector approximating the group response curve (GRC) given by:

$$G_j = \sum_{i=1}^{N} P_{ij}(\hat{\theta}) \; / \; N, \quad j = 1,-----n, \quad \text{and} \quad i= 1,----N$$

$$\text{Where} \quad P_{ij}(\hat{\theta}) = \hat{c}_j + \frac{1 - \hat{c}_j}{1 + \exp(-D\hat{a}_j(\hat{\theta}-\hat{b}_j)}$$

D=1.702 and $\hat{a}_j$, $\hat{b}_j$, and $\hat{c}_j$ are item parameter estimates.

The second extended caution index, ECI2, measures the extent to which the observed item response vector, Y_i, conforms to the group response vector G compared to the counterpart of the theoretically derived examinee response vector P_i and G. The fourth extended caution index, ECI4, is a measure that indicates the extent to which the observed vector Y_i conforms to the examinee response curve P_i compared to the group response curve vector of G to P_i (Tatsuoka & Harnish, 1983).

However, the effectiveness of unstandardized extended caution indices was found to be related to examinee ability level. Therefore, Tatsuoka & Tatsuoka (1982b) standardized the extended caution indices by subtracting their expected values and then dividing by their standard errors. These indices are denoted by ECIZ1, ECIZ2, ECIZ3, ECIZ4, ECIZ5, and ECIZ6. The second and the

fourth standardised extended caution indices can be computed relatively easily. Let θ_i denote the one-, two- or three-parameter logistic maximum likelihood estimate of θ for the i^{th} person in the test norming sample of N examinees, and let $P_{ij}(\theta_j)$ be the probability of a correct response to item j by this i^{th} examinee. ECIZ2 and ECIZ4 are then defined as follows.

$$\text{ECIZ2} = \frac{\sum \{ P_i(\hat{\theta}) - U_i \} \{ G_j - \bar{G} \}}{\{ \sum P_i(\hat{\theta})\, Q_i(\hat{\theta})\, (G_j - \bar{G})^2 \}^{1/2}}$$

where $\bar{P} = \sum P_i(\hat{\theta}) / n$, $\quad i=1----------n$

$$\text{and ECIZ4} = \frac{\sum \{ P_i(\hat{\theta}) - U_i \} \{ P_i(\hat{\theta}) - \bar{P} \}}{\{ \sum P_i(\hat{\theta})\, Q_i(\hat{\theta})\, [P_i(\hat{\theta}) - \bar{P}]^2 \}^{1/2}}$$

where $G_j = \sum_{i=1}^{N} P_i(\hat{\theta}) / N$, $\quad i = 1---------N$

and $\bar{G} = \sum_{j=1}^{n} G_j / n$, $\quad j = 1---------n$

Maximum Likelihood Indices. Levine and Rubin (1979) used an entirely different approach from the covariance approach represented by the extended caution indices. They proposed an index

denoted by Lo, which is closely related to the likelihood function , L. The likelihood function of the N examinees on an n-item test is given by:

$$L = \prod_{i=1}^{N} \prod_{j=1}^{n} P_{ij}(\hat{\theta})^{u_{ij}} (1 - P_{ij}(\hat{\theta})^{(1 - u_{ij})}$$

If a particular examinee (denoted by θ_i) does not contribute much to the maximising likelihood function , L , then it is likely that examinee i is not an appropriate representative of the group whose abilities are measured by the test. Thus, Levine and Rubin (1979) defined the appropriateness index, Lo, as:

$$Lo = \Sigma \left[u_{ij} \text{ Ln } P_{ij}(\hat{\theta}) + (1 - u_{ij}) \text{ Ln } Q_{ij}(\hat{\theta}) \right]$$

Where $P_{ij}(\hat{\theta})$ is the probability of examinee i answering item j correctly, u_{ij} is the observed item responses (0 for wrong , 1 for right) and $Q_{ij}(\hat{\theta}) = 1 - P_{ij}(\hat{\theta})$.

With Lo, aberrance of an individual's pattern of item responses is indicated by a relatively low maximum of the function θ. An atypical examinee pattern is expected to have a relatively low likelihood function because it is not likely that high ability people miss easy items or low ability people pass hard items. However, Drasgow (1982) sugested that Lo can be improved by computing Lo for the items answered, and then calculating the geometric mean likelihood, Lg, which is defined as follows;

$$Lg = \exp(Lo/n)$$

Levine and Drasgow (1983) found the Lo index to be less effective with examinees with high omit rates. For example, a low value of Lo obtained by an examinee who omitted several items is less indicative of aberrance than a higher value of Lo achieved by an examinee who omitted fewer and different items. However, Drasgow, Levine, and Williams (1985) have shown that the ''standardization process'' is useful in accounting for item omitting as well as controlling for the confounding effect of ability and appropriateness. They transformed the Lo index into a standard normal distribution and denoted it as LZ. Some reseachers denote it as Z3 or L3. The LZ index is defined as follows:

$$LZ = \frac{Lo - E(Lo)}{(var(Lo))}$$

where $E(Lo) = \Sigma \{P_{ij}(\hat{\theta}) \text{ Ln } P_{ij}(\hat{\theta}) + (1 - P_{ij}(\hat{\theta})) \text{ Ln } (1 - P_{ij}(\hat{\theta}))\}$

and $Var(Lo) = \Sigma P_{ij}(\hat{\theta})(1 - P_{ij}(\hat{\theta})\{ \text{Ln } [P_{ij}(\hat{\theta})/(1 - P_{ij}(\hat{\theta})]\}^2$

However, Hoijtink (1986) using the Rasch model found that the distribution of LZ is affected by item difficulty and examinee ability distributions. Molenaar and Hoijtink (1990) reported that the distribution of LZ is also influenced by examinee ability distributions. Thus, replications of these two studies under the two- or three-parameter model is recommended.

Levine and Rubin (1979) also suggested two other appropriateness indices based on the Gaussian model. In the Gaussian model , the ability level θ_i is allowed to vary from item to item. For each item , a new ability, θ , is sampled. These θs are assumed to be independently sampled from a normal distribution with mean θ_o and variance σ^2_g. To estimate examinee parameters using the Gaussian model , Levine and Drasgow (1982b) first estimated ICCs for a standard logistic model using LOGIST. Then they wrote the likelihood of a vector of item responses as a function of θ_o, σ^2_g, and the individual θ_i. The θ_i can be integrated out of the likelihood equation , leaving

$$Ln = \log \text{prob} \ (\ U/\theta_o, \ \sigma^2_g \ )$$

Finally, θ_o and σ_g can be estimated using numerical methods. The first appropriatenes index using the Gaussian model is the logarithm of the likelihood ratio comparing a Gaussian model to its standard model. Levine and Rubin (1979) denote it as LR. However, it will be denoted as LK in this study. It is defined as:

$$LK = Ln - Lo$$

A large LK indicates a relatively unimportant amount of variation in the ability estimates (Levine & Drasgow, 1982).

The second index obtained from the Gaussian model is the estimate, σ_g, the standard deviation of the θ_i distribution. This

index should be near zero when the major determinant of a vector of responses is a single ability. If other factors influence the examinee's responses (e.g cheating, misunderstanding directions etc), then θ_i should be greater than zero (Hulin & Parsons, 1983).

Fit Statistics. Wright (1977) discussed two statistics that can be used to evaluate whether the observed number of correct responses agrees with the number predicted from the Rasch model parameter estimates. The first, U1 , is an unweighted total fit mean square:

$$U1 = \sum_{j=1}^{n} [(u_{ij} - P_i(\hat{\theta}))^2 / P_i(\hat{\theta})(1 - P_i(\hat{\theta}))] / n$$

Where i indexes the examinee, j indexes the n-items, $P_i(\hat{\theta})$ is the probability of a correct response predicted by the Rasch model, and u_{ij} is the observed item response.

The second, W1 , is the weighted total fit mean square given by;

$$W1 = \sum_{j=1}^{n} [u_{ij} - P_i(\hat{\theta})]^2 / \sum_{j=1}^{n} P_i(\hat{\theta})[1 - P_i(\hat{\theta})]$$

The three parameter generalizations of Wright's (1977) unweigthed and weigthed fit statistics are derived by simply replacing the probability of a correct response predicted by the one parameter Rasch model with the probability of a correct response predicted by the three parameter Birnbaum model (Rudner,

1983). The unweigthed and weighted 3-parameter fit statistics are denoted as U3 and W3 respectively. Some researchers denote them as F1 and F2 respectively.

The two person fit statistics which are computed for each individual by summing over items, are differentially sensitve to different kinds of items. The unweigthed fit statistics are influenced more by very hard and very easy items, while the weighted fit statistics are more sensitive to items of near mean difficulty (Rudner, 1983). However, additional research is required to establish this fact.

Other non-standardized IRT indices have been utilized. Drasgow, Levine, and McLaughlin (1987) used two indices; the jacknife (JK) and the observed / expected curvature (O/E) which are based on the assumption that the likelihood function will have a flattened curve near its maximum indicating a poor fit.

More recently Drasgow, Levine, and McLaughlin (1987) and Drasgow and Levine (1986) have proposed the concept of an optimal index which is considered to be a theoretical index such that no other index 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 as $LR = P_{A(u)} / P_{N(u)}$ where u is the response pattern. However, they 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.

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. IRT indices also have the advantage over non-IRT indices in that the distributions of the standardized indices including LZ, ECIZ2, and ECIZ4 are approximately normal (Drasgow, Levine, & McLaughlin, 1987). However, large sample sizes are required to compute IRT based indices. A lot of computer time is also required. The indices may not be indicants of aberrance with speeded tests where omit rates may be high. However, a polychotomous model (Zh) which takes the omitted responses into account has been developed (Drasgow, Levine, & Williams, 1985). In the next chapter, a review of the literature is presented. Studies which have used IRT or a mixture of IRT and group dependent indices are reviewed.

CHAPTER II

REVIEW OF THE LITERATURE AND RESEARCH QUESTIONS

The goal of appropriateness measurement is to identify examinees with inappropriate scores solely from their response patterns. This is done in two steps. First, a general psychometric model is fitted to a large sample of nominally normal examinees. Then an index of goodnes of fit or appropriateness index is used to measure the degree to which each individual examinee's response pattern fits the model used to characterise normal behavior.

Most of the research studies to be reviewed in this chapter are those in which the effectivenes and the distributional characteristics of the appropriateness indices were investigated. To evaluate the effectiveness of an appropriateness index, the receiver operating characteristic (ROC) curve is used. To construct and use the ROC curve, a cut-off or the criterion value of the index is specified. At each criterion value, say t, the proportion of aberrant examinees correctly identified and the proportion of normal examinees improperly identified as aberrant are y(t) and x(t) respectively; then an ROC curve is obtained by plotting the [x(t), y(t)] pairs for various values of t. The x(t) is the false alarm rate and y(t) is the hit rate. An ROC curve that indicates good detection of aberrance is one that rises sharply from the origin to the upper left corner of the plot. A random classification system would produce an ROC curve that lies on a 45

degree diagonal.

In the first large scale, systematic appropriateness measurement study, Levine and Rubin (1979) investigated the effectiveness of three appropriatness indices (1) the square root of the maximum likelihood estimate of the ability variance (σ), (2) conditional marginal probability (Lo) and (3) the likelihood ratio (LK) in detecting examinees's aberrant response patterns to a multiple choice test. They used Hambleton and Rovinelli's (1973) programs to simulate SAT data. To simulate a ''normal'' candidate, first an ability θ was sampled from a normal, zero mean, unit variance population. Then the item scores for the examinee were simulated as a sequence of independent Bernoulli trials. The success probability on the i^{th} trial, $P_i(\theta)$, was computed using a three-parameter logistic model.

Examinees with varying degrees of aberrance were generated by modifying the item scores of examinees who are assumed to have responded to the items normally. To simulate a spuriously high examinee (cheating) score, k% of examinee's item responses were randomly sampled without replacement and rescored as correct, but left unchanged if correct. Four percent, 10%, 20% and 40% of the examineees' scores were generated to be spuriously high.

To generate a spuriously low score, a response vector of a 'normal responding examinee' was generated and k% of his/her items were randomly sampled without replacement and rescored as correct with a probability of 0.2 and as incorrect with a probability of 0.8 (i.e replaced with random responses). Response vectors with 4%,

10%, 20%, and 40% spuriously low scores were generated.

The three indices of appropriateness measurement, Lo, LK, and σ, were computed for the two types of aberrance. Receiving operating characteristic (ROC) curves were drawn for each index. The 20 percent spuriously low aberrance was well detected by Lo. However, Lo index did slightly better than chance for 4 percent spuriously low aberrant scores. Similar results were reported for the likelihood ratio (LK) and the σ indices.

Spuriously high aberrant candidates could be more easily detected than the spuriously low aberrant candidates. This could be attributed to the fact that more scores were changed from incorrect to correct in the spuriously high treatment and fewer scores were changed from correct to incorrect in the spuriously low treatment. However, the 40% spuriously high aberrant candidates were well detected by all three indices. The detection rates increased with the proportion of item responses changed. At 0.01 false alarm rate, Lo, LK, and σ could detect 70%, 80%, and 40% of the 40 percent spuriously high aberrant candidates.

However, Levine and Rubin's (1979) study was limited to simulated data conforming to the three- parameter logistic model. Furthermore, their use of simulated item parameters (rather than estimated item parameters) left open the question of how well appropriateness measurement would perform in applications requiring parameter estimation.

Levine and Drasgow (1982) extended the work of Levine and Rubin (1979) to more realistic conditions by using actual and

simulated data. They investigated the effects of including aberrant examinees in the norming sample and the existense of errors in estimating item parameters on the effectiveness of Lo and LK indices in detecting aberrant response patterns.

They simulated data by first sampling an ability θ from a normally distributed population with a mean of zero and unit variance. Next a number was sampled from a uniform distribution in the unit interval. If the sampled number was less than or equal to the three-parameter probability vector ($P_i(\theta)$), then the item was scored as correct; otherwise it was scored as incorrect. A total of 3200 response vectors were created to simulate the Scholastic Aptitude Test Verbal item parameters. A ''normal''sample of 2800 response vectors was created from the first 2800 response vectors in the original 3200 response vectors. Another sample labelled ''LOW 200'' was created from the response vectors between 3001 to 3200 from the original response vectors. These response vectors were modified to be 20% spuriously low by sampling 20% of the items (without replacement) and rescoring them as incorrect with a probability of 0.8 and as correct with a probability of 0.2. A ''LOW 102'' sample was created by selecting from the LOW 200 sample the 102 response vectors where at least 10% of the simulated examinees' original responses were changed in the spuriously low modification.

Their first study compared the effectiveness of Lo in detecting aberrant response patterns with estimated and with actual (simulated) item parameters. First, item parameter estimates were

estimated by LOGIST for the NORMAL 2800 response vectors. Lo was then computed for each of the NORMAL 2800 examinees by evaluating the likelihood function at $\theta = \hat{\theta}$, the LOGIST maximum-likelihood estimate of ability. The estimated item parameters were then used to compute Lo for the LOW 102 response vectors in two steps. First, maximum-likelihood ability estimates were obtained by running LOGIST for the LOW 102 data with all item parameters fixed at the values obtained from the NORMAL 2800. Secondly, Lo was calculated for each of the LOW 102 examinees by substituting θ with $\hat{\theta}$.

The results of study 1 showed very close agreement between values of Lo computed with simulation parameters and estimated parameters. However, there was a tendency for estimated parameter index values to be slightly smaller than the simulated index values for most of the aberrant examinees in the LOW 102 sample. Even closer agreement was observed for the NORMAL 2800 sample. Therefore, using estimated item parameters to compute Lo did not seem to degrade the detection rates of Lo. However, the norming and classification samples were the same. It is possible, then, that the high detection rates of Lo in this study could in part be attributed to the overfitting of estimated parameters. If another sample were used, much lower detection rates could have been observed.

In their second study, Levine and Drasgow (1982) investigated the effects of unidentified aberrants in the norming sample on the effectiveness of Lo in the aberrant response patterns. The NORMAL 2800 and the LOW 200 samples were merged to form a data file with

a large proportion of aberrant examinees. Item parameters were estimated using the 3000 simulated examinees' data. Lo was next computed for all examinees by evaluating the likelihood function at the LOGIST maximum estimate of ability. New index values for the LOW 102 sample were compared with the exact parameter (parameters used to simulate data) index values. Again, the results showed that estimating item parameters in large samples and with a large proportion of spuriously low examinees (1:14) did not noticeably degrade appropriateness measurement.

Levine and Drasgow (1982) used actual data in their third and fourth studies. In their third study, they investigated the effects of using overlapping norming and classification samples on the detection rates of Lo and LK. Three thousand ''low omitting'' examinees were sampled from the ETS Scholastic Aptitude Test-Verbal. LOGIST was used to estimate item parameters from these 3000 nominally normal examinees. A file of 200 aberrant examinees was then created by applying 20 percent spuriously low modifications to answer sheets for examinees 2801, 2802.....3000. Lo and LK were computed for all the aberrant examinees.

The responses to the Graduate Record Examination-Verbal section by 10000 examinees were used to investigate the effects of using distinct norming and classification samples in study 4. A norming sample of 3000 examinees was created by selecting 1,2,3,11,12,13.....,9991,9992,9993 examinees. A classification sample of 2470 examinees was created by selecting those examinees who had answered as many as 86 of the 95 GRE-V items and the

response vectors of 200 of them were subjected to the 20 percent spuriously low modifications. The remaining 2270 response vectors formed the normal group. Lo was computed for the 200 aberrant and 2270 normal response vectors.

Lo and LK were found to have high detection rates at low false alarm rates in study 3. However, this exceptional performance of LK and Lo could be attributed to the overfitting of item parameters since the norming and the classification samples overlapped. In study 4, Lo attained high hit rates of 10, 20, and 40 percent at false alarm rates of 3%, 14%, and 37% respectively.

These four studies show that some appropriateness measurement techniques may be robust to errors in estimation of item parameters and to the inclusion of unidentified aberrant response patterns in the test norming sample. However, like Levine and Rubin (1979), Levine and Drasgow (1982) only considered examinees who answered all or nearly all items and ignored which wrong answer was chosen. The classification and the norming samples were drawn from the same populations. It is possible that different results could have been observed if the norming and the classification samples were drawn from different populations. The effects of test length, IRT model, item difficulty, and examinee ability distributions on the effectiveness and distributions of LK and Lo is not considered in the study.

Rudner (1983) evaluated nine different indices of appropriateness measurement. They included C, r, NCI, MCI, W1, U3, W3, LZ, and rb (point biserial correlation). To generate data,

item parameters were specified first, then examinees' abilities and their responses to the test items were simulated. Birnbaum's three-parameter logistic model was used to relate item and examinee characteristics to item responses. Two independent data sets were generated and denoted as T1 and T2. The data set, T1, was designed to simulate the performance of examinees on a high quality commercial examination based on Lord's (1968) parametarization of 80 verbal Scholastic Aptitude Test items. The data set, T2, was designed to simulate item responses to a shorter teacher made classroom test and was based on Bejar, Weiss, and Kingsbury's (1977) parameterization of 45 items for a general Biology examination. Mean discrimination, difficulty, and pseudo-guessing item parameters were 1.07, 0.58, and 0.16 respectively for T1 and 1.09, 0.08, and 0.25 for T2 respectively.

After the three item parameters were specified, an examinee's ability was generated by randomly drawing an ability parameter from a normally distributed population with a mean of zero and standard deviation of one. The item parameters and the examinee's ability determined the probability of a correct response to the item for the examinee. The probability of a correct response was then transformed to an observable right-or-wrong response by comparing it to a uniformly distributed random number between 0 and 1. A response was coded correct ($u_{ij} = 1$) when its associated probability was greater than the random number and incorrect ($u_{ij} = 0$)when it was less.

Spuriously low total scores were obtained by randomly

selecting k% of the examinees's item responses and changing them to incorrect responses. Spuriously high total scores were obtained by changing k% of incorrect responses to correct.

Item responses for T1 were simulated for a control group of 2000 (non-aberrant) examinees and eight experimental groups of 100 normally distributed examinees with altered item responses. Item responses for the shorter T2 test were simulated for a control group of 2000 examinees and six experimental groups of 100 examinees each. Spuriously high scores were created with 5, 10, 15, and 20 item responses for T1 and with 5, 10, and 15 item responses for T2 changed from incorrect to correct. The same number of item responses of T1 and T2 were changed from correct to incorrect to simulate spuriously low scores. Then the nine appropriateness indices were computed. Correlations among the nine indices were computed for T1 and T2.

Rudner found r, NCI, and W1 values to be highly intercorrelated while U3, W3, LZ and U1 values were less correlated. Also with a critical value of 0.05 (Type I error), the effectiveness of all the nine indices in detecting aberrance increased as a function of the number of altered item responses. W1 and rb did not accurately identify Test 1 examinees with spuriously high scores.

W3 tended to identify the greatest number of examinees with spuriously high or spuriously low scores in both T1 and T2 data sets whereas LZ worked well in identifying examinees with spuriously high and spuriously low scores on Test 1 and with

examinees with spuriously low scores in Test 2. W1 followed a similar trend but did not do well with spuriously high scores on Test 1. U1, r, and rb tended to work well on the shorter test (T2) but poorly on the longer test (T1). U3 worked well with T1 but worked poorly with T2 data sets.

However, Rudner's (1983) study leaves a lot of questions unanswered. He does not clearly indicate the variables which were manipulated. If test type was the variable of study; then the attained results can be attributed to the confounding effect between test type and test length. Scores were also simulated to be normally distributed, but could the same results be observed if the scores were simulated to be non-normal?. Therefore, we cannot generalize his findings to other situations and settings.

Birenbaum (1985) compared nine IRT-based indices of appropriateness measurement with respect to the total test scores. The indices evaluated in the study were ECI1, ECI2, ECI4, ECIZ1, ECIZ2, ECIZ4, LO, LZ, and U1. The sample consisted of 280 tenth graders who took a 1982 multiple choice test in reading comprehension in English as a Second Language in the greater Tel-Aviv area of Israel. Another 30 responses were simulated. A student's response was termed as inappropriate if the student did not write his/her name on the answer sheet, otherwise it was termed appropriate. Out of 280 students selected 238 wrote their names and 42 did not. The simulated responses (N= 30) were also considered to be appropriate. Item parameters based on the two-parameter logistic model were calculated using LOGIST . All the nine indices

were computed for each examinee.

Standardized and unstandardized extended caution indices along with LZ were found to be highly intercorrelated. The unstandardized appropriateness index Lo, and the unweghted fit statistic, U1, yielded low intercorrelations with other indices. In general, the nine indices examined yielded low, almost negligible linear relations with the total test score.

The above findings suggest that appropriateness measures (index values) can be considered as a source of nonredundant information that can be added to the information derived from the total test score. However, ECI1, ECI2, ECI4, ECIZ1, Lo, and U2 showed some curvilinear trend with respect to the total test score; indicating that when applied can result in inflated scores at the extremes of the ability scale. Apart from being unrelated to the total test score, ECIZ2, ECIZ4, and LZ detected 74%, 74%, and 75% of the inappropriate examinee patterns respectively at a false alarm rate of 25%. However, we cannot generalize the results of this study to other situations and settings because the author defined an aberrant examinee as one who did not write his/her name on the answer sheet; it is possible that some examinees termed as aberrant in the study may have had non-aberrant response patterns but forgot to write his/her name. Other variables such as test length, item difficulty distribution, examinee ability distribution, etc were not manipulated in the study.

Drasgow, Levine, and Williams (1985) examined the distributions of Lo, LZ, and Zh under different conditions of

parameter estimation and with the presence of errors in the norming sample. They also investigated the effectiveness of LZ and Zh indices in detecting aberrant responses of examinees with moderately high non-response rates.

They used 75000 response vectors of examinees who wrote the April, 1975 Scholastic Aptitude Test-Verbal Section. A norming sample of 3000 response vectors was created by selecting every 20^{th} examinee begining with the first examinee and scoring their responses as correct, incorrect, omitted and not reached. LOGIST was used to estimate item and person parameters according to the three-parameter model. These item parameter estimates were then used to construct histograms summarizing the pattern of option selection at various ability levels for the 49470 examinees following the first 25000 examinees in the data set. They were then sorted into 25 ability categories according to their ability estimates. The 4^{th}, 8^{th}, 12^{th}, ...96^{th} percentile points from the standard normal distribution were used as cutting scores to form the ability categories. The frequencies of option selection were also determined for each of the 85 SAT-V items in each ability category; and then converted into proportions to give $P_{vk}(\theta)$ values of the histogram model. $P_{vk}(\theta)$ is the proportion of examinees with ability θ choosing option v of item k.

To examine the distribution of Lo, LZ, and Zh, ability estimates were computed for the first 500 response vectors with unrestricted omit rates using the three-parameter model. Lo, LZ, and Zh were calculated for a total of 464 of them because their

ability estimates were in the interval $-2.05 \leq \theta \leq 2.05$.

In the results section, the values of Lo were found to change as a function of the examinees' ability levels. However, the values of LZ and Zh were relatively constant over examinees' ability levels. In fact examinees with extreme ability estimates had near zero (mean of a normal distribution) values of LZ and Zh. However, high omitting rates of easy items and not finishing the examination were found to cause Zh to indicate aberrance. Therefore, examinees who omitted more than 35 percent of the test items or reached less than 77 percent of the test items were excluded from subsequent analysis.

To investigate further the distributions of LZ and Zh with restricted omitting, a sample of 3478 nominally normal examinees were formed from examinees 10001 to 14000 who had ability estimates in the interval $-2.05 \leq \theta \leq 2.05$. LZ and Zh were then computed for these 3478 nominally normal examinees.

The cumulative frequency distributions of LZ and Zh showed LZ and Zh to have asymetric distributions. There were relatively few examinees with index scores between -2.0 and 0.0 and relatively many examinees with index scores between 0.0 and 1.2. The two indices also had distributions which were significantly different from the normal distribution ($\alpha = 0.01$) by using the Kolmogorov-Smirnov test.

Using simulated data the authors replicated the preceding study. A sample of 4000 response vectors, sorted into 25 ability categories was generated using the three-parameter logistic model

and the three-parameter histogram model. Hypothetical probabilities of correct responses on dichotomously scored items and the hypothetical probabilities of option selection on polychotomously scored items were computed. For each simulated examinee, an ability was sampled from the standard normal distribution truncated to the interval $-2.05 \leq \theta \leq 2.05$. Responses to 85 items were then simulated as 85 independent multinomials with response probabilities obtained by substituting sampled ability in the three-parameter logistic ICCs and histogram option response functions. LZ and Zh appropriateness indices were computed using estimated abilities.

The conditional distributions of Zh were found to be less invariant than the conditional distributions of LZ. Further, the cumulative proportions of LZ were approximately similar to the proportions found with real data (SAT-V).

To investigate the effectiveness of LZ and Zh in detecting aberrant response patterns, a sample of 3478 nominally normal, low omitting examinees was used as a normal group. An aberrant group of examinees was formed by considering only those who omitted less than 35% and reached not less than 77% of the test items. Five ability categories were formed: low [-2.05, -0.80], moderately low [-0.80, -0.24], average [-0.24, 0.24], moderately high [0.24, 0.80], and high [0.80, 2.05]. Starting with examinee 10001 on the SAT-V tape, 300 examinees were selected for each of the five ability categories and subjected to spurious modifications. To simulate spurious high scores, 10%, 20%, and 30% of the examinee's

original responses were randomly selected without replacement and rescored as correct. Ten, 20, and 30 percent spuriously low scores were created by first determining the proportion of omitted items (q), and then selecting (without replacement) 10, 20, and 30 percent of the examinee's original responses and scoring them as omitted with a probability of q. Options A to E were selected with a probability of (1-q)/5. The three-parameter model was used to re-estimate the ability of examinees whose responses were modified before computing LZ and Zh appropriateness indices.

ROC curves were used to display the effectiveness of an index in detecting simulated aberrance. Detectability of spuriously low response vectors by both LZ and Zh appropriateness indices increased as a function of ability. This could be attributed to the fact that spuriously low treatment changed relatively few responses of low ability examinees and changed relatively more responses of high ability examinees. In contrast, the detectability of spuriously high modifications decreased as a function of ability. However, LZ performed better than Zh in detecting spuriously high responses and Zh performed better than LZ in detecting spuriously low responses. The detection rates of both indices in spuriously low and high modifications increased with the number of items modified.

The most important finding of this study is that LZ can have as high detection rates of aberrant response patterns as its polychotomous counterpart, Zh. Additional research to determine the distributional characteristics of this index (LZ) under different

experimental and testing conditions seems warranted. For example, little is known about its distribution under varying conditions of item difficulty and examinee ability distributions.

In a follow-up study, Drasgow and Levine (1986) evaluated the effectiveness of LZ and Zh with respect to the optimal index (LR) in detecting spuriously high and spuriously low response vectors in an 85-item test. To study the upper limit on the detectability of mildly aberrant response vectors, a normal sample of 4000 response vectors was generated using the three-parameter model. The ability parameters were randomly sampled from a normal (0,1) distribution truncated to the interval $-2.05 \leq \theta \leq 2.05$. A 10% spuriously high treatment sample was created by first sampling 2000 normal response vectors and randomly selecting nine items and rescoring them as correct. Another 2000 response vectors were generated and subjected to the 10% spuriously low treatment by randomly selecting nine items and rescoring them as incorrect with a probability of 0.8 and as correct with a probability of 0.2.

The ability and item parameter estimates were calculated using LOGIST according to the three-parameter model. The ability estimates obtained were then used to compute LZ and LR appropriateness scores for the three samples.

The detection rates of LZ appropriateness index were found to be very close to the detection rates of LR index with LZ being about 90% of the detection rates of LR at corresponding false alarm rates. At a 5% false alarm rate, LR detected 24% of the spuriously high response patterns. The detection rates of LZ were larger in

the spuriously high conditions than in the spuriously low conditions. At low false alarm rates LZ detected as many as 65% of the spuriously low scores as did the optimal index.

Drasgow and Levine (1986) also compared the detection rates of LR and LZ in dichotomously and polychotomously scored item responses. The polychotomous responses consisted of five-option multiple choice items with omitting allowed. The three-parameter logistic model was used to estimate the abilities of 49470 examinees from the April, 1975 administration of the SAT-V (Drasgow, 1985). Twenty five ability groups were formed by using the 4^{th}, 8^{th}, -------96^{th} percentiles points from the normal distributions. The frequency of each option selection in each ability category was determined. The skipped and not reached items were treated as one category. Probabilities of option responses were computed by linear interpolations between ability category medians (i.e; the 2^{nd}, 6^{th},------98^{th} percentile points of the normal distributions).

A norming sample of four thousand normal response patterns was generated. Two thousand normal response patterns were also generated and subjected to 10% spuriously high treatment by randomly selecting nine items (without replacement) and rescoring them as correct. Another 2000 normal response patterns were generated and subjected to 10% spuriously high treatment by randomly selecting nine items (without replacement) and rescoring them as correct with a probability of 0.2 and as incorrect with a probability of 0.8. In each case, the abilities were randomly

sampled from the normal (0,1) distribution truncated to the interval $-2.05 \leq \theta \leq 2.05$. The responses were scored dichotomously and polychotomously and LZ, optimal LZ, Zh, and optimal Zh indices computed for the 10% spurious modifications.

As in the Drasgow et al.(1985) study, LZ was found to have much higher detection rates of spuriously high response vectors than the Zh index. At low false alarm rates, the Zh index yielded detection rates which were only about one-fourth as high as the rates of its optimal index. In contrast, the detection rates for LZ were about 65% of the detection rates of its optimal index. In the 10% spuriously low treatment, Zh optimal index detected 41% of the aberrant response vectors at a 5% false alarm rate. Dichotomous scoring reduced the detectability of spuriously low response vectors. However, polychotomous item response models are less tractable than their dichotomous counterparts. Furthermore, there are only a few examples of substantial gains from polychotomous analysis. Therefore, LZ seems to be a promising practical appropriatenes index to be used in practical testing situations; hence additional research to determine the effectiveness and the distributional characteristics of this index under different conditions of testing is required.

Hoijtink (1986) investigated the robustness of the distributions of Lo, LZ, ECI2, ECI4, ECIZ2, ECIZ4, U1, and W1 against variations in the distributions of item difficulty and examinee ability using the Rasch model. Simulated item and ability parameters were used.

The simulation design included three types of distributions for the item difficulty parameters, each consisting of twenty items with item difficulty of -3 for the easiest item and +3 for the most difficult item. In the uniform distribution of item difficulty, item parameters were equally spaced. In the normal distribution, the distance between item difficulties increased from the centre to the extremes while in the right skew distribution, the distance between item difficulties increased from left to right.

Four ability distributions were used: Normal, uniform, right skew, and left skew. A total of 2500 response patterns were simulated for each combination of item difficulty and the four ability distributions. The eight appropriateness indices were computed for each sample. For each appropriateness index the 90^{th}, 95^{th}, and 99^{th} percentiles were computed along with their 95% confidence intevals. The mean, standard deviation, skewness, and kurtosis of each index were also computed.

An index was said to be robust if its 90^{th}, 95^{th}, and 99^{th} percentile estimates remained the same under different combinations of item difficulty and ability distributions. An answer pattern was classified as aberrant, if in a sample that was simulated to fit the Rasch model, it belonged to the 5% or 1% patterns with the worst fit.

All the eight indices had similar results. It was found that the 95^{th} percentile estimates (only results of 95^{th} percentiles were reported) of all the indices were not robust against variations in item difficulty and ability distributions. The overlap between

confidence intervals was often very small and sometimes even absent which indicated that the differences between the 95th percentile estimates were not caused by sampling variations but by the variations of the distributions of the appropriatenes indices with the total score.

Therefore, the 95th percentile estimate is not the 95th percentile estimate for each ability group, but a weighted mean of the 95th percentile estimates of the ability groups. These results suggest that if a sample consists of many medium ability persons, the 95th percentile estimates will decrease. If many low and high ability persons are in the sample, the estimate will increase. This is an undesirable feature for a Rasch model based person fit index; aberrance becomes sample dependent in the sense that given a set of items, a given answer pattern can be aberrant in one and non-aberrant in another sample. This means that score patterns can only be aberrant within the context of the ability group to which they belong.

However, these results are only generalizable to situations where the data fit the one-parameter model. Future research should investigate the effects of item difficulty and examinee ability distributions on the distributions and effectivenes of appropriateness indices under the two- or three-parameter models.

Molenaar and Hoijtink (1990), using the Rasch model, investigated how the the marginal distribution of Lo changes when the distribution of ability remains normal, but its mean and variance are altered. In particular, they examined the effects of

ability distribution on the probability of exceedance of four most aberrant response patterns (0111, 0001, 0101, 0011) in a four-item test with item difficulties of -2.0, -1.0, -0.5 and 1.5. The probability of exceedance was defined as the probability of obtaining a response pattern given the ability distribution. If the probability of exceedance was less than a certain prespecified value (e.g 0.001), then the response pattern was said to be aberrant.

The effects of the mean on the probability of exceedance was examined by using normal samples with ability means of -3.0, -2.0, -1.0, 0, 1.0, 2.0, and 3.0 and each with ability variance of +1. To investigate the effects of ability variance, normal samples with mean abilities of zero and ability variances of 0.25, 0.5, 1.0, 2.0, and 4.0 were used. In both cases, Lo was calculated for all the possible response patterns. These response patterns were next ordered according to their Lo values. The probability of occurence of each response pattern was calculated. The probabilities of exceedance (cumulative probabilities) were also computed for the four most aberrant response patterns.

The results showed that the probability of exceedance of the four most aberrant response patterns changed as a function of one of the two parameters (mean or variance) when the other was kept constant. Thus, this study indicates that the ability distribution of the number correct scores affects the distributions of Lo appropriateness index. To extend this study, future researcher's should investigate the effects of varying the mean and the variance

of the examinees' ability distributions on the distributional characteristics of other apropriateness indices using one-, two-, and three-parameter models.

Drasgow, Levine, and McLaughlin (1987) examined the effectiveness of nine apropriateness indices by comparing them to the optimal index (LR). The indices investigated were LZ, U3, W3, C, ECIZ2, ECIZ4, JK, O/E, and IOV (item-option-variance). The three-parameter model was used to estimate the abilities of 49470 examinees from the 85-item April 1975 administration of SAT-V. Item responses with five options were simulated using the Levine and Drasgow (1985) histogram. Examinees were sorted into 25 ability categories by using the 4^{th}, 8^{th},----------96^{th} percentile points of the normal (0,1) distribution as the cutting scores. The proportions of examinees choosing each option were computed for each of the 25 ability groups. Probabilities of option selection were then computed by linear interpolation between category medians at the 2^{nd}, 6^{th}, ----------98^{th} percentile points.

Five samples of normal response patterns were created by first sampling 3000 normal response patterns in the normal (0,1) distributions truncated to the interval $-2.05 \leq \theta \leq 2.05$. Low (-2.05 to -1.05), moderately low (-0.70 to -0.55), average (-0.55 to -0.05), moderately high (0.55 to 0.70), and high (1.49 to 2.05) θ samples of N=200 each were formed.

Polychotomous response vectors were then generated for each ability level. For each item, the associated histogram was used to compute the conditional probabilities of the six possible responses

(treating non-response as a sixth response). A number was sampled from the uniform distribution in the unit interval and a simulated response was obtained by determining where the random number was located in the cumulative distribution corresponding to the conditional probabilities. The nine appropriateness indices were then computed for each of the five samples.

The ROC curves showed that IOV, C, JK, O/E, ECIZ2, and U1 indices were poorly standardized. However, the standardization of LZ, U3, W3, and ECIZ4 was shown to be relatively good across all the ability groups.

To determine the effectiveness of the nine indices under different conditions of aberrance, a norming sample of 3000 response vectors was created by sampling 3000 responses from the normal (0,1) distribution truncated to the interval $-2.05 \leq \theta \leq 2.05$. A normal sample of 4000 response vectors was also generated. Two thousand aberrant response vectors were created for each condition by varying three factors: Type of aberrance (spuriously high or low), severity of aberrance (mild, moderate) and level of ability (very low, low, low average, high average, high, and very high).

Spuriously high response patterns were created by first generating normal response vectors; then k% of the simulated responses were randomly sampled without replacement and rescored as correct. Spuriously low response patterns were also created by first generating normal response vectors and then randomly selecting k% of them and replacing them with random responses.

Mildly aberrant and moderately aberrant response patterns were created by using k = 15% and k = 30% respectively. All the nine indices were computed for each examinee in the normal, spuriously high, and spuriously low samples.

The detection rates of the nine indices decreased as a function of ability in the spuriously high treatment samples. LR, LZ, ECIZ2, and ECIZ4 indices could detect 89%, 75%, 81%, and 79% respectively in the 30% spuriously high treatment samples with low θs at 1% false alarm rate. At a 1% false alarm rate LR, LZ, W3, ECIZ2, and ECIZ4 could also detect 78%, 51%, 53%, 51%, and 57% respectively of the 30% spuriously high treatment samples with low average θs (31^{st} through 48^{th} percentiles).

The effectiveness of these indices in detecting spuriously low treatment samples increased as a function of θ. At 1% error rate LR had a 47% detection rate for the 15% spuriously low treatment and 79% for the 30% spuriously low treatment of aberrant samples with average θs. LZ the next best index, could detect 35% of the 30% spuriously low treatment samples and ECIZ4 could detect 28%.

Further, the practical appropriateness indices had detection rates that were closer to the rates of LR for the 15% and the 30% spuriously low samples with θs in the 65^{th} through 92^{nd} percentiles. At a 1% error rate the detection rates of LR, LZ, W3, ECIZ2, and ECIZ4 were 91%, 71%, 62%, 64%, and 63% respectively. For the spuriously low sample with high θs (percentiles 93^{rd} and above), the detection rates of these indices were 97%, 95%, 91%, 94%, and 90% respectively at a 1% error rate.

Levine and Candell (1990) examined the asymptotic performance of LR, LZ, U3, W3, and ECIZ4 using computer adaptive tests (CAT). They simulated fixed length CAT administration by using ten ability levels (corresponding to 5^{th}, 15^{th},.....85^{th}, and 95^{th} percentiles of a normal distribution) to generate response vectors for 15-, 20-, and 25-item CATs. In the non-aberrant conditions, dichotomous responses were determined for each item using the three-parameter logistic function. In the aberrance conditions for the 15-item CAT, the probability of a correct response to the initial 1, 2, 3, 4, and 5 responses was set at 0.2; the remaining responses were determined by the three-parameter logistic function. In the aberrance conditions for the 20- and 25-item CATs, the probability of a correct response to each of the first 5 items was set at 0.20. For each test length and aberrance condition, mean ability estimates and mean posterior standard deviation (psd) were calculated over the 1000 simulees at each θ level.

They reported that the mean ability estimates of the above average θs with no aberrant condition were underestimated. For instance, an examinee at the 95^{th} percentile giving misinformative responses to the first five items would receive on average, an ability estimate of the 25^{th} percentile. Similar results were reported for the 20- and the 25-item CATs.

They used the 15-item test to determine the detection rates of the five appropriateness indices. Detection rates of LR were evaluated by using random responses to the initial 1, 2, 3, 4, and 5 items. The non-optimal indices were evaluated for random

responses to the initial 5 items.

For each appropriateness index, they enumerated all the possible response patterns for a 15-item test. The items that the deterministic CAT would administer were identified and the appropriateness index calculated for the pattern. The probabilities of observing the pattern under conditions of aberrance and non-aberrance were computed. Next, all patterns were ordered from the largest to the smallest on the magnitude of the appropriateness index. Beginning with the first pattern (i.e the largest index value), and continuing with each successive pattern, cumulative pattern probabilities for aberrance and non-aberrance were calculated to give the 'hit' rate and the 'false alarm' rate respectively for each index. Theoretical ROC curves were also constructed for the five indices.

The results showed that LR could detect all the five (k =1, 2, 3, 4 and 5) initial patterns of random responses to the 15-item test at a 1% error rate. When aberrance was defined as random responses to the initial five items, LR index could correctly identify 49% of the truly aberrant response patterns at a false alarm rate of 0.001 and achieved a detection rate of 85% at a false alarm rate of 0.10. LZ, U3, W3 and ECIZ4 could detect 29%, 27%, 32% and 27% of the aberrant response patterns (random responses to the initial 5 items) at a 10% error rate respectively. At all false alarm rates, W3 had the highest detection rates of aberrant response patterns. LZ was the next best performing non-optimal index.

It is evident from this study that non-optimal indices are ineffective in detecting aberrant responses in CAT. However, these results are not generalizable because several factors (e.g test length, item difficulty, and examinee ability, IRT model etc) were not manipulated.

Noonan (1990) investigated the effects of test length (40 and 80 items) and IRT model (two- and three-parameter models) on the distributions of LZ, W3, and ECIZ4 indices in nonaberrant response patterns. He also investigated the effects of test length, IRT model, type of aberrance (spuriously low and spuriously high), and level of aberrance (10%, 15%, and 30%) on the effectiveness of the three indices. Data were generated by computer in four combinations of two test lengths and two IRT models to produce non-aberrant response patterns. Aberrant responses were also generated in twenty four combinations of two test lengths, two IRT-models, two types of aberrance and three levels of aberrance. The data generated had no omit rates.

Non-aberrant response vectors for 40 and 80-item test under two- and three- parameter model were generated. Item difficulty was selected from a range of -2.00 to 2.00 from a uniform distribution for the two test lengths using each of the two models. Item discrimination parameters were also selected from a uniform distribution from 0.40 to 1.50; and guessing parameters were selected from a range of 0.05 to 0.20 for the three-parameter and set to zero for the two-parameter model.

To examine the distribution of these indices, Datagen (1985)

was used to generate 2000 response vectors for each of the four combinations of two test lengths and two IRT models. At each replication, the mean, standard deviation, kurtosis, and skewness were computed for each index. The values of ECIZ4 and W3 indices at 99^{th}, 95^{th}, 90^{th}, and 75^{th} and the values of LZ at 1^{st}, 5^{th}, 10^{th}, and 25^{th} percentile points were also computed (since small negative values of LZ indicate aberrance). A total of 50 replications were used in each combination.

The results showed that ECIZ4 was least affected by test length and IRT model and its distribution was approximately normal (0,1). For LZ, test length had an effect for all cutoff scores. LZ also produced an approximately normal (0,1) distribution. W3 was significantly affected by test length and IRT model at all cutoff scores.

To examine the detection rates of LZ, ECIZ4, and W3 appropriateness indices in detecting aberrant response patterns 10%, 20%, and 30% spuriously low and spuriously high response vectors were also generated for a total of 4000 examinees in each of the twenty four combinations. Spuriously high scores were produced by randomly selecting 10%, 15%, or 30% (without replacement) of the incorrect items and rescoring them as correct. Spuriously low scores were obtained by sampling 10%, 15%, or 30% of the correct items and rescoring them as incorrect. The ability of each examinee was re-estimated and then the three indices computed for each examinee (response vector) for each of the twenty four combinations.

LZ was found to produce the highest detection rates at a 1% false alarm rate with lower levels of aberrance while ECIZ4 generally produced the highest detection rates at 30% level of aberrance. The 80-item test produced higher detection rates than the 40-item test. The two-parameter model also produced higher detection rates than the three-parameter model. In summary, LZ and ECIZ4 appropriateness indices provided consistent information regardless of testing conditions.

However, these results are only limited to situations where the item difficulties are uniformly distributed and the examinee ability distribution is normally distributed. In fact Hoijtink (1986) using the Rasch model showed that the disributional characteristics of these two indices (and others) are affected by the item difficulty and examinee ability distributions. Therefore, additional research should determine the effectivenes and the distributional characteristics of these two indices under different conditions of item difficulty and examinee ability distributions under a two- or a three- parameter model.

Summary and the Research Questions.

The review of the literature has highlighted numerous issues pertaining to appropriateness measurement in both non-aberrant and aberrant response patterns. For instance, additional research is required to determine the distributional characteristics of some of the appropriateness indices under different experimental and testing conditions. It is also evident from the review of the

literature that some indices are well 'standardized' while some are poorly 'standardized'. Among the well standardized appropriatenes indices are LZ, ECIZ4, W3, and ECIZ2 .

However, Birenbaum (1985) reported the W3 appropriateness index to have a curvilinear relationship with the total test score indicating that the W3 index yields inflated values at the extremes of the ability continuum. Noonan (1990) also reported the W3 appropriateness index to have non-normal distributions under different conditions of test length and IRT model. In constrast, LZ and ECIZ4 appropriateness indices have been reported to have approximately normal distributions and to be least related, linearly or curvilinearly, to the total test scores, indicating that they provide non-redundant information (Drasgow et al., 1985, 1986; Noonan, 1990).

Thus, additional research is required to determine the effectiveness and distributions of these two indices under different experimental and practical testing conditions. In particular, the effects of item difficulty and examinee ability distributions on the effectiveness and the distributional characteristics of LZ and ECIZ4 using a two- or three-parameter model should be investigated.

Therefore, the purpose of this study is to determine the effects of item difficulty and examinee ability distributions on the distributional charateristics of LZ and ECIZ4 appropriateness indices using a three-parameter model. The effectiveness of these two indices under different combinations of item difficulty and

examinee ability distributions, type of aberrance, and level of aberrance will also be investigated.

The research questions are:

1. What are the effects of item difficulty and examinee ability distributions on the distributions of LZ and ECIZ4 appropriateness indices in known non-aberrant response patterns?.
2. To what extent are LZ and ECIZ4 appropriateness indices effective in detecting aberrant response patterns under varying conditions of item difficulty and examinee ability distributions, type of aberrance and level of aberrance?.

CHAPTER III

METHODOLOGY

This chapter presents the design which was used in this research to examine two questions related to LZ and ECIZ4 appropriateness indices. Simulated data and not real data were used. Data were generated according to the three-parameter model to simulate the responses of examinees to 60 multiple-choice items using Datagen, a fortran computer program developed by Hambleton and Rovinelli (1977) and modified by Carlson (1985). In previous research, the three-parameter logistic model has been found to be adequate for modelling the multiple choice items on the Scholastic Aptitude Test-Verbal section (Drasgow, 1982; Levine & Rubin, 1979; Levine & Drasgow, 1982; Rudner, 1983; Drasgow et al.,1985, 1986, 1987), Graduate Record examination-Verbal Section (Drasgow, 1982; Levine & Drasgow, 1982) and simulated data (Noonan, 1990; Candell Levine, 1990). The LOGIST computer program (Wood, Wingersky, & Lord, 1976) was used to estimate item parameters.

To evaluate the effectiveness of appropriateness indices, most researchers have used the design devised by Levine and Rubin (1979). In this design, a study begins with the test norming sample that consists of N examinees' responses (either real or simulated) to n items. Item parameters for a test model are estimated using the test norming sample. These item parameter estimates are then used to estimate examinee's ability and to compute appropriateness

indices. A similar design was employed in this study and a FORTRAN77 program written by Drasgow (1985) was used to compute LZ and ECIZ4 scores.

In this study, the effects of item difficulty and examinee ability distributions on the distributional characteristics and effectiveness of LZ and ECIZ4 appropriateness indices were investigated. Hoijtink (1986), using the Rasch model, reported that examinee ability and item difficulty distribution affect the distributions of appropriateness indices. Molenaar and Hoijtink (1990), using the Rasch model, reported that examinees ability distributions affect the distributions of Lo.

Three distributions of item difficulty were used. These distributions were those which are usually found in real life situations and they were generated to simulate the distributions of items typical of diagnostics tests (items used to identify students who need remedial courses), power (placement) tests, and certification and licensing tests. Items typical of diagnostic tests were generated to have uniform distributions in the interval $-3.0 \leq b \leq +1.2$. These test items were expected to provide maximum information (differentiate) at the low ability range. Item difficulties typical of those found with power tests were generated to have a uniform distribution in the interval $-3.0 \leq b \leq +3.0$. These items were expected to provide equal information (differentiate) over the ability range (Van der Flier, 1982). Item difficulties typical of those found with certification and licensing examinations were generated in such a manner that they

would provide maximum information at the high ability range. They were generated to have a uniform distribution in the interval $-1.2 \leq b \leq +3.0$. In summary, all three distributions of item difficulties were generated to have uniform distributions. Uniform distribution of item difficulties is what to be expected for most tests.

Since the objective of this study was to investigate the effects of item difficulty and ability distributions and not item discrimination or the guessing parameters, uniform distributions of item discrimination and guessing parameters were used for each replication. In all the replications, the discrimination parameters were generated in a such a manner that $0.60 \leq a \leq 1.50$ and to have uniform distributions. The guessing parameters were generated in such a manner that $0.05 \leq c \leq 0.20$ and to have uniform distributions. Such distributions of guessing parameters are typical of five-option multiple choice tests.

Three distributions of ability were considered. In each replication, a normal distribution of examinees' ability with a standard deviation of 0.6 but with different means were used. Molenaar et al. (1990) in a simulation study found that the distributions of appropriateness indices were affected by the position of the mean and the standard deviation of the examinees' ability distribution even when the examinees' ability distribution remained normal. The ability distributions used were those typical of low, medium, and high ability examinees. Thetas typical of low ability examinees were generated to have normal distributions with

a mean of -1.2 with a standard deviation of 0.6. Medium ability thetas typical of medium ability examinees were generated to have a normal distribution with a mean of zero and a standard deviation of 0.6. High ability thetas typical of high ability examinees were generated to have a normal distribution with a mean of +1.2 and a standard deviation of 0.6.

To examine the effects of item difficulty and examinee ability distributions on the distributional characteristics of LZ and ECIZ4 appropriateness indices, data were generated in nine combinations of item difficulty and examinee ability distributions. In each replication, data were generated to simulate the responses of 2000 examinees to 60 test items according to the three-parameter model. LOGIST (Wood, Wingersky, & Lord, 1976) was used to estimate item parameters. LZ and ECIZ4 appropriateness indices were computed for each examinee in each of the nine combinations of item difficulty and examinee ability distributions.

The mean, standard deviation, skewness, and kurtosis were computed for each index in each of the nine combinations of item difficulty and examinee ability distributions. The values of ECIZ4 at the 99^{th}, 95^{th}, 90^{th}, and 75^{th} percentile points and the values of LZ at the 1^{st}, 5^{th}, 10^{th}, and 25^{th} percentile points were also computed. A total of 50 replications were used in each combination.

The means and standard deviations were computed over the 50 replications for the mean, standard deviation, kurtosis, skewness, and the four percentiles of each index. These statistics were used to test the robustness of LZ and ECIZ4 indices under the varying

conditions of item difficulty and examinee ability distributions.

To examine the main and the interaction effects of item difficulty and examinee ability distributions, 3 x 3 MANOVAs were performed using the mean, skewness, standard deviation, and kurtosis values for each index over the 50 replications as the dependent variables and item difficulty and examinee ability as the independent variables. To further examine the efffects of item difficulty and examinee ability distributions, 3 x 3 MANOVAs were performed also using the means of the four percentiles of each index as the dependent variables.

To examine the effectiveness of LZ and ECIZ4 appropriatenes indices in detecting aberrant response patterns under different combinations of item difficulty and examinee ability distributions, type of aberrance, and level of aberrance, response vectors were generated using Datagen. Two samples each consisting of 500 normal response vectors (one for spuriouly low and one for spuriously high modifications) were also generated in each of the nine combinations and subjected to spurious treatment. An examinee with a spuriously high test score was simulated by selecting 20% or 10% of the examinee's original responses without replacement and changing incorrect answers to correct, but they were left unchanged if correct. An examinee with a spuriously low test score was simulated by first randomly selecting 20% or 10% of the examinee's original responses without replacement and changing correct responses to incorrect, but they were left unchanged if incorrect.

LZ and ECIZ4 appropriateness indices were then computed for

the aberrant response vectors. The effectiveness of each index was evaluated by examining the extent to which it separated normal and aberrant response vectors solely on the basis of appropriateness index scores. The percentile estimates obtained for each index at each false positive rate were used as cutoff scores. The results of this study and the discussion of them are presented in the next chapter.

CHAPTER IV

RESULTS AND DISCUSSION

In this chapter, the results of this study and their discussion are presented. First, the effects of item difficulty and ability distributions on the distributional characteristics of LZ and ECIZ4 indices in non-aberrant response patterns are presented. Secondly the effectiveness of the indices in aberrant response patterns are presented. The chapter concludes with the summary of the results.

Distribution of the Indices.

An index with a known statistical distribution is a useful tool for practical purposes. Such an index can be used to estimate population parameters with a certain degree of confidence. With respect to appropriateness indices, researchers such as Tatsuoka and Tatsuoka (1982a, 1982b), Drasgow, Levine and Williams (1985), Hoijtink and Molennar (1990), Hoijtink (1986), and Noonan (1990) have investigated the distributions of the appropriateness indicess.

The mean, standard deviation, skewness, and kurtosis over 50 replications were used to investigate the distributional characteristics of the two indices under different combinations of item difficulty and ability distributions. Three uniform distributions of item difficulty parameters were used: low (-3.0

to 1.2), medium (-3.0 to 3.0) and high (1.2 to 3.0). Normal distributions of ability with means of -1.2, 0.0 and 1.2 and each with a standard deviation of 0.6 were designated as low, medium and high ability distributions respectively.

To examine the effects of item difficulty and ability ditributions on the descriptive statistics of LZ and ECIZ4, separate multivariate analysis of variance were performed on the mean, standard deviation, skewness, and kurtosis of each index over the 50 replications as the dependent variables and item difficulty and ability distributions as the independent variables. Using Pillai's statistics at 0.05 level of significance, the results showed that the combined dependent variables (the mean, standard deviation, skewness and kurtosis) of LZ were significantly affected by ability distributions, $F(8, 876)=341.6$, $P<0.001$, item difficulty, $F(8, 876)=429.4$, $P<0.001$, and by their interactions, $F(16, 1339)=129.5$, $P<0.001$. It was also found that the mean, standard deviation, skewness and kurtosis of ECIZ4 were significantly affected by the main effects of ability distributions, $F(8, 876)=468.5$, $P<0.001$, item difficulty, $F(8, 876)=213.9$, $P<0.001$, and by their interactions, $F(16, 1339)=248.0$, $P<0.001$.

To determine which dependent variables contributed to the overall significance of the main effects of ability distributions, item difficulty and their interactions, univariate analyses of variance were also conducted. Again the ANOVA results showed that all the dependent variables of LZ and ECIZ4 contributed to the

overall multivariate significance of the main effects of ability distribution, item difficulty, and their interactions at 0.01 level of significance. Univariate ANOVAs were tested at 0.01 level of significance so that the overall experimentwise error rate would be held at less than 0.05.

Following significant interactions, simple effects analysis of variance, followed by Scheffe post hoc analysis were also conducted in order to investigate how the dependent variables differed over ability groups in each item difficulty category and how they differed over item difficulty parameters in each ability distribution. Table 1 contains the mean values of LZ and ECIZ4 over 50 replications and the Scheffe post hoc results.

The values of the mean of the LZ index were found to be very close to the expected value under all the nine combinations of item difficulty and ability distributions. However, under the conditions of low item difficulty parameters, a mean value less than the expected value was obtained when the ability distributions were low and a value bigger than the expected value was obtained when the ability distributions were high. In fact the values of the mean increased as a function of ability under the low item difficulty. The opposite trend was exhibited under the high item difficulty. Under the high item difficulty, a value of the mean less than the expected value was obtained when the ability distributions were high and a mean value bigger than the expected value was obtained when the ability distributions were low. However, the values of the mean of LZ were relatively similar under

Table 1

The Means of LZ and ECIZ4 over 50 Replications and the Scheffe Post Hoc Results.

	Ability Groups			
Item. diff.	Low (1)	Medium (2)	High (3)	Scheffe test
LZ				
Low (1).	-0.001	0.000	0.008	1&3, 2&3
Med. (2).	-0.002	-0.003	-0.002	none
High (3).	0.007	-0.001	-0.001	1&2, 2&3
Scheffe test	1&3, 2&3	1&2	all	
ECIZ4.				
Low (1).	-0.008	-0.005	0.226	1&3, 2&3
Med. (2).	-0.014	-0.008	-0.008	1&3, 1&2
High (3).	-0.045	-0.008	-0.005	1&3, 1&2
Scheffe test	all	1&2, 1&3	1&2, 1&3	

the medium item difficulty parameters.

Under all the three item difficulty parameters, ECIZ4 had the smallest values of the mean when the ability distributions were low; and it had the highest values of the mean when the ability distributions were high. The highest value of the mean (0.226) was obtained under the combination of low item difficulty and high ability distributions; and the smallest value (-0.045) of the mean was obtained under the combinations of high item difficulty and low ability distributions. The values of the means of ECIZ4 increased as a function of ability distributions under high item difficulty parameters. They also increased as a function of ability distributions under the low item difficulty. Overall, the mean values of LZ were found to be less affected by item difficulty and ability distributions than the mean values of ECIZ4.

Table 2 includes the standard deviation values of LZ and ECIZ4 indices over 50 replications and the Scheffe post hoc results. The mean standard deviations of LZ were found to be different from the expected value of 1.0. Under low item difficulty parameters, the standard deviation of LZ deviated most (0.751) from the expected value when the ability distributions were high and deviated least (0.988) when the ability distributions were low. In fact the mean values of the standard deviation decreased as a function of ability distributions under the low item difficulty parameters. In the case of high item difficulty parameters, the mean values of the standard deviation deviated most (0.972) from the expected value when the ability distributions were low, and it deviated least (1.004) when

Table 2

The Standard Deviations of LZ and ECIZ4 Over 50 Replications and the Scheffe Post Hoc Results.

	Ability Groups			
Item. diff.	Low (1)	Medium (2)	High (3)	Scheffe test (3)
Low (1).	0.988	0.926	0.751	all
Med. (2).	0.989	0.974	0.988	1&2, 2&3
High (3).	0.972	1.004	0.979	1&2, 2&3
Scheffe test	1&3, 2&3	all	1&2, 1&3	
ECIZ4				
Low (1).	0.941	0.945	0.778	1&3, 2&3
Med. (2).	0.827	0.949	0.908	all
High (3).	0.742	0.880	0.962	all
Scheffe test	all	1&3, 2&3	all	

the ability distributions were medium.

Under the low item difficulty parameters, the standard deviation of ECIZ4 deviated most (0.778) under high ability distributions and deviated least when the ability distributions were medium (0.945) or low (0.941). Under high item difficulty parameters, the standard deviation was much smaller (0.742) than the expected value when the ability distributions were low; and it was closer to the expected value (0.962) when the ability distributions were high. Similar results were obtained with the medium item difficulty parameters. In general, the mean values of the standard deviations of LZ were less affected by item difficulty and ability distributions than the mean values of the standard deviations of ECIZ4.

Table 3 contains the skewness values of LZ and ECIZ4 over 50 replications and the Scheffe post hoc results. The mean skewness values showed the distributions of LZ and ECIZ4 indices to be somewhat skewed. LZ index had negative mean skewness values which ranged from -0.648 to -0.285 for all combinations of item difficulty and ability distributions suggesting that LZ is skewed to the left. Further, the mean skewness values of LZ were found to increase as a function of item difficulty; and they decreased as a function of ability distributions. ECIZ4 had positive mean skewness values which ranged from 0.015 to 0.299 for all combinations of item difficulty and ability distributions. This suggests that ECIZ4 is positively skewed. The mean skewness values of ECIZ4 increased as a function of ability distributions with high item difficulty,

Table 3.

The Skewness Values of LZ and ECIZ4 Over 50 Replications and the Scheffe Post Hoc Results.

	Ability Groups			
Item. diff.	Low (1)	Medium (2)	High (3)	Scheffe test.
LZ				
Low (1).	-0.311	-0.398	-0.648	all
Med. (2).	-0.338	-0.394	-0.566	all
High (3).	-0.285	-0.326	-0.410	all
Scheffe test	2&3	1&3, 2&3	all	
ECIZ4				
Low (1).	0.267	0.266	0.015	1&3, 2&3
Med. (2).	0.052	0.050	0.058	none
High (3).	0.123	0.299	0.284	1&3, 2&3
Scheffe test	1&3, 2&3	1&2	1&2, 1&3	

but decreased as a function of ability distributions with low item difficulty. Overall, the mean skewness values of ECIZ4 were less affected by item difficulty and ability distributions than the mean skewness values of LZ.

Table 4 shows the kurtosis values of LZ and ECIZ4 and the Scheffe post hoc results. The mean kurtosis values of LZ and ECIZ4 indices exhibited slightly different patterns from that exhibited by the mean skewness values. The mean kurtosis values of LZ increased as a function of ability distributions and decreased as a function of item difficulty parameters. They ranged from 0.038 to 0.907. However, the mean kurtosis values of ECIZ4 increased as a function of item difficulty parameters and they decreased as a function of ability distributions under the high item difficulty parameters. They ranged from 0.067 to 0.577. The mean kurtosis values of ECIZ4 were found to be less affected by item difficulty and ability distributions than the mean kurtosis values of LZ.

The Scheffe post hoc results showed that the standard deviation and the skewness values of LZ significantly differed among all the three ability groups under the low item difficulty. But the mean values of LZ were not significantly different among the three ability groups under the medium item difficulty. The kurtosis values of the low ability group significantly differed with the kurtosis values of both the high and the medium ability groups under all the three item difficulty categories. However, the standard deviation values of ECIZ4 differed significantly among the three ability groups under the medium and under the high item

Table 4.

The Kurtosis Values of LZ and ECIZ4 Over 50 Replications and the Scheffe Post Hoc Results.

	Ability Groups			
Item. Diff.	Low (1)	Medium (2)	High (3)	Scheffe test
LZ				
Low. (1)	0.081	0.212	0.907	1&3, 2&3
Med. (2)	0.165	0.147	0.271	1&3, 2&3
High (3)	0.038	0.074	0.224	1&3, 2&3
Scheffe test	none	1&3, 2&3	all	
ECIZ4.				
Low. (1)	0.098	0.067	0.097	none
Med. (2)	0.196	0.102	0.141	none
High.(3)	0.577	0.212	0.108	all
Scheffe test	1&3, 2&3	1&3, 2&3	none	

difficulty. The kurtosis and the skewness values of ECIZ4 were not significantly different among all the ability groups under the medium item difficulty.

In summary, the means, standard deviations, skewness, and kurtosis of LZ and ECIZ4 were not robust against the variations of item difficulty and ability distributions. The values of the means of LZ were generally closer to the expected values than the mean values of ECIZ4 for all combinations of item difficulty and ability distributions. It was also found that the mean standard deviations of LZ were closer to the expected values than the mean standard deviations of ECIZ4 index. The absolute kurtosis values of LZ were consistently higher than the absolute kurtosis values of ECIZ4 suggesting that the distribution of LZ is slightly more peaked than the distribution of ECIZ4. The results also showed LZ to have negative skewness values while ECIZ4 had positive skewness values; indicating that LZ is skewed to the left and ECIZ4 is skewed to the right. That is, each index is skewed towards the direction of it's aberrance.

Further, the mean and standard deviation values of the two indices were found to be very different from their expected values when the ability parameters did not match the item difficulty parameters. For instance, under a combination of low item difficulty and high ability distributions or under a combination of high item difficulty and low ability distributions, the smallest or the highest values of the two statistics were observed. In situations where item difficulty and ability distributions were

matched, the values of the mean and the standard deviation were very close to their expected values. Under these combinations, the skewness and the kurtosis values were also found to be very low. Therefore, the distributions of LZ and ECIZ4 indices approximate a normal distribution when the item difficulty and ability distributions are matched.

This finding is consistent with previous research findings. Drasgow, Levine, and Williams (1985), using a normal (0, 1) distribution of ability and a wide range of item difficulty parameters found the distributions of LZ to approximate that of a normal distribution. Noonan (1990), also using a normal (0, 1) distribution of ability parameters and a wide range of item difficulty found the distributions of LZ and ECIZ4 to aproximate that of a normal distribution.

Determining the Cutoff Scores.

Since only the tails containing the fit values of aberrant response patterns are usually of interest, the 99^{th}, 95^{th}, 90^{th} and the 75^{th} percentile estimates were computed for ECIZ4. The 1^{st}, 5^{th}, 10^{th} and the 25^{th} percentile estimates were computed for LZ because small values of LZ are indicative of aberrance. These four percentile estimates in each of the nine combinations of item difficulty and ability distributions would be used as cutoff points to classify response patterns as aberrant or non-aberrant. An answer pattern would be classified as aberrant if it belonged to the worst 1, 5, 10 or 25 percent of the patterns with the worst fit

in a sample that was simulated to fit a three parameter model.

To examine the effects of item difficulty and ability parameters on the percentile estimates of LZ and ECIZ4, separate MANOVAs were conducted on the four percentile estimates of each index as the dependent variables and item difficulty and ability distributions as the indepedent variables. Using Pillai's statistics at 0.05 level of significance, the results showed that the combined dependent variables (four percentile estimates) of LZ were significantly affected by ability distributions, $F(8, 876)=129.5$, $P<0.001$, item difficulty, $F(8, 876)=200.8$, $P<0.001$, and by their interactions, $F(16, 1339)=86.69$, $P<0.001$. The main effects of ability distributions, $F(8, 876)=139.5$, $P<0.001$, item difficulty, $F(8, 876)=60.1$, $P<0.001$, and their interactions, $F(16, 1339)=99.0$, $P<0.001$ significantly affected the four percentile estimates of ECIZ4.

To determine which dependent variables contributed to the overall significance of the main effects of ability distributions, item difficulty and their interactions, univariate analyses of variance were also conducted. The ANOVA results showed that all the percentile estimates for each index contributed to the overall multivariate significance of the main effects of ability distribution, item difficulty, and their interaction.

Following significant interactions, simple effects analysis of variance followed by Scheffe post hoc analysis were also conducted in order to investigate how the four percentile estimates for each index differed over ability groups in each item difficulty level

and how they differed over item difficulty categories in each ability distribution. Tables 5, 6, 7, and 8 present the 1st, 5th, 10th and 25th percentile estimates of LZ and the 99th, 95th, 90th, and 75th percentile estimate of ECIZ4 respectively over 50 replications and the Scheffe post hoc results.

The results showed that all the four percentile estimates of LZ to be different from the expected values. The mean values of the 1st and the 5th percentile estimates of LZ were smaller than their expected values of -2.33 and -1.65 respectively for most combinations of item difficulty and ability distributions and they increased as a function of ability distributions under the low item difficulty parameters. The 1st and the 5th percentile estimates of LZ tended to be underestimated while the 10th and the 25th percentile estimates tended to be overestimated.

The four percentile estimates of ECIZ4 were found to be different from the expected values. Except for very few cases, the 95th, 90th and 75th percentile estimates of ECIZ4 were less than the expected values 0f 1.65, 1.29 and 0.68 respectively. The 99th percentile estimates did not show any pattern in terms of it's magnitude. The results also showed that the percentile estimates of ECIZ4 deviated most when the item difficulty parameters did not match the ability distributions. For example, very low percentile estimates were observed under combinations of low item difficulty and high ability distributions and under combinations of high item difficulty and low ability distributions, suggesting that item difficulty and ability distributions have some impact on the

Table 5.

The 1st and the 99th Percentile Estimates of LZ and ECIZ4 Over 50 Replications and the Scheffe Post Hoc Results.

	Ability Groups			
Item. diff.	Low (1)	Medium (2)	High (3)	Scheffe test
LZ				
Low. (1)	-2.497	-2.422	-2.137	all
Med. (2)	-2.556	-2.542	-2.734	1&3, 2&3
High.(3)	-2.442	-2.580	-2.556	1&2, 1&3
Scheffe test	1&2, 2&3	1&2, 1&3	all	
ECIZ4.				
Low. (1)	2.365	2.388	2.079	1&3, 2&3
Med. (2)	2.114	2.428	2.315	all
High.(3)	1.817	2.261	2.442	all
Scheffe test	all	1&3, 2&3	all	

Table 6.

The 5th and the 95th Percentile Estimates of LZ and ECIZ4 Over 50 Replications and the Scheffe Post Hoc Results.

	Ability Groups			
Item. diff.	Low (1)	Medium (2)	High (3).	Scheffe test
LZ.				
Low. (1)	-1.716	-1.632	-1.362	all
Med. (2)	-1.726	-1.700	-1.767	1&3, 2&3
High.(3)	-1.666	-1.741	-1.722	1&2, 1&3
Scheffe test	1&3, 2&3	all	all	
ECIZ4.				
Low. (1)	1.616	1.620	1.472	1&3, 2&3
Med. (2)	1.412	1.637	1.566	all
High.(3)	1.201	1.508	1.652	all
Scheffe test	all	1&3, 2&3	all	

Table 7.

The 10th and the 90th Percentile Estimates of LZ and ECIZ4 0ver 50 Replications and the Scheffe Post Hoc Results.

	Ability Groups			
Item. diff.	Low (1)	Medium (2)	High (3)	Scheffe test
LZ				
Low. (1)	-1.300	-1.225	-0.989	all
Med. (2)	-1.308	-1.291	-1.139	1&3, 2&3
High.(3)	-1.275	-1.323	-1.300	all
Scheffe test	none	all	all	
ECIZ4				
Low. (1)	1.227	1.230	1.185	none
Med. (2)	1.064	1.240	1.183	all
High.(3)	0.894	1.136	1.257	all
Scheffe test	all	1&3, 2&3	1&3, 2&3	

Table 8.

The 25th and the 75th Percentile Estimates of LZ and ECIZ4 Over 50 Replications and the Scheffe Post Hoc Results.

	Ability groups			
Item. diff.	Low (1)	Medium (2)	High (3)	Scheffe test
LZ.				
Low. (1)	-0.640	-0.587	-0.428	all
Med. (2)	-0.640	-0.629	-0.610	1&3, 2&3
High.(3)	-0.624	-0.652	-0.620	1&2, 2&3
Scheffe test	1&3, 2&3	all	all	
ECIZ4.				
Low. (1)	0.599	0.611	0.763	1&3, 2&3
Med. (2)	0.515	0.605	0.578	all
High.(3)	0.425	0.554	0.619	all
Scheffe test	all	1&3, 2&3	1&2, 1&3	

percentile estimates. However, the percentile estimates of LZ and ECIZ4 were were found to be very close to the expected values when the ability estimates matched the item difficulty parameters.

The Scheffe post hoc results further showed that the 1st, 5th, 10th and 25th percentile estimates of LZ significantly differed among all the three ability groups under the low item difficulty and they significantly differed between the low and the high and between the medium and the high ability groups under the medium item difficulty. They also significantly differed between the low and the medium and between the low and the high ability groups under high item difficulty, an indication that the percentile estimates of LZ are more stable under the medium and under high item difficulty than they are under the low item difficulty.

However, the 99th, 95th, 90th, and the 75th percentile eatimates of ECIZ4 significantly differed among all the three ability groups under the medium and under the high item difficulty. The 99th, 95th, and 75th percentile estimates of ECIZ4 significantly differed between the low and the high and between the medium and the high ability groups under the low item difficulty. This suggests that the percentile estimates of ECIZ4 are more stable under the low item difficulty than they are under the medium and under high item difficulty.

In summary, the percentile estimates of LZ and ECIZ4 indices were found to be sensitive to the variations of item difficulty and ability distributions. The four percentile estimates obtained for each index were found to be different from the expected values in

all the nine combinations. This finding is not surprising considering the fact that the descriptive statistics (distribution) of the two indices varied with the changes in the side conditions of item difficulty and ability distributions.

The four percentile estimates of LZ were found to be closer to the expected values than the percentile estimates of ECIZ4. This could be attributed to the fact that LZ approximated a normal distribution better than ECIZ4. For LZ, the marginal mean values of the 1^{st}, 5^{th}, 10^{th} and 25^{th} percentile estimates were -2.396, -1.660, -1.350 and -0.614 respectively. The 99^{th}, 95^{th}, 90^{th} and 75^{th} marginals of the percentile estimates of ECIZ4 were found to be 2.244, 1.520, 1.157 and 0.585 respectively.

However, the four percentile estimates for each index were found to be different from combination to combination. The main effects of item difficulty and ability distributions and their interactions were found to be significant at 0.05 level. Further analysis showed that all the dependent variables (the four percentile estimates for each index) contributed to the overall multivariate significance.

Discussion.

Several researchers have strongly maintained that appropriateness indices which have a known distribution are useful for practical purposes. However, there are numerous factors which affect testing in practical testing situations. The distribution of an effective appropriateness index should be relatively invariant across different testing conditions. Examinee sample size is one

constraint which varies from time to time. Different testing conditions require different parameter estimations. However, Noonan (1990) using 2- and 3- parameter models found that the distribution of LZ and ECIZ4 indices were not seriously affected by item parameter model and test length.

Hoijtink (1986) using a Rasch model found the distribution of all the indices studied to be sensitive to the side conditions of item difficulty parameters and ability distributions. The percentile estimates were also found to be non-robust against the variations of item difficulty parameters and ability distributions.

In the present study, the descriptive statistics and the percentile estimates of LZ and ECIZ4 were also found to be very sensitive to the side conditions of item difficulty and ability distributions using a three parameter model. The mean values of the mean and the standard deviation were found to vary as a function of ability distributions and item difficulty parameters. The mean and the standard deviation values of the two indices over fifty replications were closer to the expected values when the item difficulty parameters matched the ability distributions.

Therefore, the distribution of LZ and ECIZ4 should be expected to approximate those of a normal distribution if a test with high item difficulty parameters is administered to a group of examinees with high ability estimates or if a test with low item difficulty parameters is administered to a group of examinees with low ability estimates. However, if a test with high item difficulty parameters is administered to a group of examinees with low ability estimates

or if a test with low item difficulty parameters is administered to a group of examinees with high ability estimates, then the distribution of the two indices should be expected to deviate most from the normal distribution.

This finding is not unique. Drasgow et al. (1985) reported the distribution of LZ index to approximate that of a normal distribution. In their study the item difficulty parameters matched the ability distributions because they used normal (0, 1) ability distributions and a wide range of item difficulty parameters. Noonan (1990) reported that the distribution of LZ and ECIZ4 were approximately normal under all combinations of test length and IRT model. He too used item difficulty parameters which matched the ability estimates.

With respect to the cutoff points, the results of this study showed that the percentile estimates of LZ and ECIZ4 were also affected by item difficulty parameters and ability distributions . However, the percentile estimates did not exhibit any particular pattern with respect to their magnitude when item difficulty parameters and ability distributions were varied.

The four percentile estimates of each index were found to be different from the expected values. The overall 3 by 3 MANOVA was shown to be significant at 0.05 level. Followup ANOVAs further showed that all the percentile estimates contributed to the significance of the overall multivariate test. With very few exceptions, Scheffe post hoc results showed that the four percentile estimates of LZ were always different among all the item

difficulty categories under high ability distributions. They were also different among all ability groups under the low item difficulty parameters. On the other hand, the four percentile estimates of ECIZ4 were always different among all item difficulty categories under the low ability distributions and they were always different among all ability groups under high item difficulty.

Given the results of the descriptive statistics, the results of the percentile estimates were not surprising. Different distributions should be expected to yield different percentile scores. It is therefore difficult to have exact cutoff scores for determining the detection rates of the two indices. The results of the present study indicated that different cutoff points should be used for each of the nine combinations of item difficulty and ability distributions.

However, it is difficult to compare the cutoff points used in this study with those used by previous researchers because different side conditions yield different cutoff scores. This is partly due to the fact that previous researchers like Drasgow et al. (1985), Birenbaum (1985), and Noonan (1990) assumed that the item difficulty parameters were uniformly distributed with a wide range. They also assumed the examinee ability parameters to have normal (0, 1) distributions. In this study the item difficulty and ability distributions were manipulated.

Detection Rates for LZ and ECIZ4 in Aberrant Response Patterns.

The strengths and weaknesses of the appropriatenes indices can be assessed via their detection rates of aberrant response patterns. The detection rates of indices are determined by examining the proportion of correct classifications of aberrant response patterns at given false alarm rates. An efficient appropriateness index should identify a large proportion of aberrant response patterns at very low false alarm rates. It's distribution should also be independent of ability level of nonaberrant response patterns.

Table 9 presents the percentage of aberrant response patterns correctly classified by LZ and ECIZ4 indices at the 0.01 false positive rate. In the 20% spuriously high aberrant response patterns, the detection rates of LZ ranged from 0% to 26% and they ranged from 0% to 7% in the 10% spuriously high aberrant response patterns. They ranged from 1% to 33% in the 20% spurioulsy low aberrant response patterns and they ranged from 0% to 20% in the 10% spuriously low aberrant response patterns. Further, under high item difficulty parameters, the percentage of aberrant response patterns correctly identified by LZ in the 20% spuriously high treatment samples decreased as a function of ability distributions. They increased as a function of ability distributions in the 20% spuriously low aberrant response patterns under the low item difficulty parameters. Spuriously low aberrant response patterns were found to be more detectable by LZ than the spuriously high aberrant response patterns at the 0.01 false positive rates.

Table 9.

The Percentage of Aberrant Response Patterns Correctly Identified by LZ and ECIZ4 at 0.01 False Positive Rate.

	20% SPURIOUSLY HIGH						20% SPURIOUSLY LOW					
	LZ			ECIZ4			LZ			ECIZ4		
	Ability distributions											
Item. diff.	Low	Med.	High	Low	Med.	High	Low	Med.	High	Low	Med	High
LoW	2	0	0	0	0	0	3	17	33	5	14	36
Med.	3	0	0	5	0	0	6	15	30	8	15	32
High.	26	16	2	23	12	3	4	1	4	2	1	2

	10% SPURIOUSLY HIGH						10% SPURIOUSLY LOW					
	LZ			ECIZ4			LZ			ECIZ4		
	Low	Med.	High	Low	Med.	High	Low	Med	High	Low	Med	High
Low	4	1	0	2	2	0	3	7	20	1	5	13
Med.	4	1	1	3	1	2	0	7	20	0	5	9
High.	7	4	2	6	4	2	0	2	4	0	2	3

The total percentage of aberrant response patterns correctly classified by ECIZ4 in the spuriously high treatment samples decreased as a function of ability distributions and increased as a function of item difficulty parameters. They ranged from 0% to 23% in the 20% spuriously high aberrant response patterns and they ranged from 0% to 6% in the 10% spuriously high aberrant response patterns. But they increased as a function of ability distributions and decreased as a function of item difficulty parameters in the spuriously low treatment samples. They ranged from 1% to 36% in the 20% spuriously low aberrant response patterns and they ranged from 0% to 13% in the 10% spuriously low aberrant response patterns. Spuriously low aberrant response patterns were also found to be more detectable by ECIZ4 than the spuriously high aberrant response patterns at 0.01 false positive rates.

Spuriously high treatment samples were more detectable by LZ than by ECIZ4 under high item difficulty and they were more detectable by ECIZ4 than by LZ under the low and under the medium item difficulty parameters. Spuriously low treatment samples were more detectable by ECIZ4 than by LZ under the low and under the medium item difficulty parameters and they were more detectable by LZ than by ECIZ4 under high item difficulty parameters.

Table 10 shows the percentage of correctly classified aberrant response patterns by LZ and ECIZ4 indices at the 0.05 false positive rates. High proportions of the spuriously high aberrant response patterns were detected by LZ when the item difficulty parameters were high and low proportions were detected under the

Table 10.

The Percentage of Aberrant Response Patterns Correctly Identified by LZ and ECIZ4 at 0.05 False Positive Rate.

	20% SPURIOUSLY HIGH						20% SPURIOUSLY LOW					
	LZ			ECIZ4			LZ			ECIZ4		
	Ability distributions											
Item diff.	Low	Med.	High	Low	Med.	High	Low	Med.	High	Low	Med.	High
Low	6	2	4	10	3	1	11	35	54	12	33	56
Med.	8	2	0	13	3	0	15	32	55	19	37	57
High.	47	36	9	43	31	9	3	6	14	8	4	12

	10% SPURIOUSLY HIGH						10% SPURIOUSLY LOW					
	LZ			ECIZ4			LZ			ECIZ4		
	Low	Med.	High	Low	Med.	High	Low	Med.	High	Low	Med.	High.
Low	11	7	3	9	6	2	8	18	40	5	15	30
Med.	16	5	7	15	8	7	5	14	31	4	12	18
High.	20	12	6	18	11	8	4	5	13	6	5	11

low and under the medium item difficulty parameters. The detection rates of LZ ranged from 0% to 47% in the 20% spuriously high aberrant response patterns and they ranged from 3% to 20% in the 10% spuriously high aberrant response patterns. In the spuriously low treatment samples, the detection rates of LZ were highest under the low item difficulty parameters and they were consistently low under the high item difficulty parameters. With the exception of very few cases, the detection rates of LZ decreased as a function of ability distributions and increased as a function of item difficulty parameters in the spuriously high treatment samples. But they increased as a function of ability distributions and decreased as a function of item difficulty parameters in the spuriously low treatment samples.

Similar trends of detections were exhibited by ECIZ4 index. In the spuriously high treatment samples, the detection rates of ECIZ4 decreased as a function of ability under high item difficulty. Under high item difficulty, the detection rates of ECIZ4 ranged from 9% to 43% for the 20% spuriously high aberrant response patterns and they ranged from 8% to 18% for the 10% spuriously high aberrant response patterns. In the spuriously low treatment samples, the detection rates of ECIZ4 were high under the low and under the medium item difficulty, and they increased as a function of ability. Under the medium item difficulty, the detection rates of ECIZ4 ranged from 19% to 57% in the 20% spuriously low aberrant response patterns and they ranged from 5% to 30% for the 10% spuriously low aberrant response patterns under the low item

difficulty.

At 0.05 false positive rates, spuriously low aberrant response patterns were more detectable by the two indices than the spuriously high aberrant response patterns. It was also observed that high proportions of the spuriously high aberrant response patterns were more detectable under the high item difficulty parameters whereas spuriously low aberrant response patterns were more detectable under the low item difficulty parameters. This could be attributed to the fact that more responses are changed from incorrect to correct and fewer are changed from correct to incorrect under the high item difficulty parameters. But more responses are changed from correct to incorrect and fewer are changed from incorrect to correct under the low item difficulty parameters. Spuriously low and spuriously high aberrant response patterns were more detectable by LZ than by ECIZ4 under the high item difficulty and they were more detectable by ECIZ4 than by LZ under the medium and under the low item difficulty parameters.

Table 11 shows the percentage of aberrant response patterns correctly classified by LZ and ECIZ4 at 0.10 false positive rate. The detection rates of LZ decreased as a function of ability distributions in the spuriously high treatment samples. Under high item difficulty, the detection rates of LZ ranged from 15% to 57% in the 20% spuriously high aberrant response patterns and they ranged from 12% to 30% in the 10% spuriously high aberrant response patterns. In the spuriously low treatment samples, the detection rates of LZ increased as a function of ability distributions. Under

Table 11.

The Percentage of Aberrant Response Patterns Correctly Identified by LZ and ECIZ4 at 0.10 False Positive Rate.

	20% SPURIOUSLY HIGH						20% SPURIOUSLY LOW					
	LZ			ECIZ4			LZ			ECIZ4		
	Ability distributions											
Item. diff.	Low	Med.	High	Low	Med.	High	Low	Med.	High	Low	Med.	High.
Low	11	4	1	16	8	4	19	48	65	21	47	64
Med.	13	3	1	21	6	2	22	47	74	26	50	69
High.	57	48	15	54	43	17	6	10	24	13	10	19

	10% SPURIOUSLY HIGH						10% SPURIOUSLY LOW					
	LZ			ECIZ4			LZ			ECIZ4		
	Low	Med.	High	Low	Med.	High	Low	Med.	High	Low	Med.	High
Low.	18	12	8	17	12	6	11	26	52	9	23	39
Med.	26	13	13	22	16	12	8	23	40	8	20	26
High.	30	21	12	28	18	13	7	9	19	12	10	17

the medium item difficulty, the detection rates of LZ ranged from 22% to 74% in the 20% spuriously low aberrant response patterns and they ranged from 11% to 52% in the 10% spuriously low aberrant response patterns under the low item difficulty. Higher proportions of spuriously low aberrant response patterns were detected by LZ than the proportions of the spuriously high aberrant response patterns.

Like LZ, the detection rates of ECIZ4 decreased as a function of ability distributions in the spuriously high aberrant response patterns and increased as a function of ability distributions in the spuriously low aberrant response patterns. Spuriously low aberrant response patterns were also more detectable by ECIZ4 than the spuriously high aberrant response patterns. At 0.10 false positive rates, LZ correctly classified higher proportions of 20% spuriously low and 20% spuriously high aberrant response patterns than ECIZ4 under the high item difficulty.

Table 12 presents the percentage of correctly identified aberrant response patterns by LZ and ECIZ4 at 0.25 false positive rate. The percentage of the spuriously high aberrant response patterns correctly classified by LZ at 0.25 false positive rate were slightly higher than the percentage of the spuriously high aberrant response patterns classified by ECIZ4 under high item difficulty. Under high item difficulty, LZ correctly classified 77%, 67%, and 34% and ECIZ4 correctly classified 71%, 63%, and 37% of the 20% spuriously high aberrant response patterns when the ability distributions were low, medium and high respectively. ECIZ4

Table 12.

The Percentage of Aberrant Response Patterns Correctly Identified by LZ and ECIZ4 at 0.25 False Positive Rate.

	20% SPURIOUSLY HIGH						20% SPURIOUSLY LOW					
	LZ			ECIZ4			LZ			ECIZ4		
	Ability distributions.											
Item diff.	Low	Med.	High	Low	Med.	High	Low	Med.	High	Low	Med.	High
Med.	27	14	2	33	21	13	38	68	83	41	65	78
Med.	31	9	3	37	22	13	40	67	87	43	69	84
High.	77	67	34	71	63	37	14	24	38	28	24	38

	10% SPURIOUSLY HIGH						10% SPURIOUSLY LOW					
	LZ			ECIZ4			LZ			ECIZ4		
	Low	Med.	High	Low	Med.	High	Low	Med.	High	Low	Med.	High
Low.	37	24	17	34	28	16	27	45	66	24	42	57
Med.	44	29	22	44	33	27	21	39	50	22	37	46
high.	55	42	25	50	39	29	17	25	38	27	27	35

had higher detection rates of the spuriously high treatment samples under the low and under the medium item difficulty. In the spuriously low aberrant response patterns, LZ had higher detection rates than ECIZ4 under the low and under the medium item difficulty.

In summary, 20% spuriously low aberrant response patterns were found to be more detectable by LZ than the 20% spuriously high aberrant response patterns. Further, the 10% spuriously low aberrant response patterns were found to be more detectable than the 10% spuriously high aberrant response patterns. Given a particular type of aberrance, aberrant response patterns in the 20% spurious samples were found to be more detectable than the detectability of aberrant response patterns in the 10% spurious treatment samples. This implies that the detection rates of aberrant response patterns by LZ increases with the level of aberrance. At low false positive rates (0.01 & 0.05), the detection rates of LZ were higher than the detection rates of ECIZ4 under the high item difficulty and the detection rates of ECIZ4 were higher than the detection rates of LZ under the low and under the medium item difficulty.

Spuriously high aberrant response patterns were more detectable under high item difficulty parameters and spuriously low aberrant response patterns were more detectable under the low item difficulty parameters. This could be attributed to the fact that more responses are changed from incorrect to correct and few responses are changed from correct to incorrect under high item

difficulty parameters and more responses are changed from correct to incorrect and few responses are changed from incorrect to correct under low item difficulty parameters. Drasgow et al. (1985, 1987), Birenbaum (1985) and Noonan (1990) reported similar results.

Further, the detection rates of the aberrant response patterns by ECIZ4 were also found to exhibit the same patterns of detection as LZ index in all the correponding experimental conditions. The 20% spuriously low aberrant response patterns were found to be more detectable by ECIZ4 than the 20% spuriously high aberrant response patterns. Higher proportions of the 10% spuriously low aberrant response patterns were also found to be more detectable than the 10% spuriously high aberrant response patterns. Aberrant response patterns in spurious high treatment samples were more detectable under high item difficulty parameters and they were more detectable under the low item difficulty parameters in the spuriously low treatment samples. The detection rates of ECIZ4 were also found to increase with the level of aberrance. Similar results were also reported by Noonan(1990), Drasgow et al. (1985), and Birenbaum (1985). The detection rates of LZ sligthly exceeded the detection rates of ECIZ4 in the spuriously high treatment samples at low false positive rates. However, the two indices had similar detection rates in the spuriously low treatment samples.

Discussion.

The results of the detection rates of LZ and ECIZ4 appropriateness indices in this study are consistent with the results reported by researchers such as Drasgow et al. (1985,

1987), Rudner (1983), Noonan (1990), and Candell and Levine (1990). In particular, the high detection rates of LZ and ECIZ4 confirm the findings of Noonan (1990). The power of the LZ index and the tendency to identify larger proportions of aberrant response patterns with spuriously high scores is also consistent with the findings of Rudner(1983), Birenbaum (1985), and Drasgow et al.(1985).

The two indices identified higher proportions of aberrant response patterns in the 20% spuriously low treatment samples than in the 20% spuriously high treatment samples. Ten percent spuriously low aberrant response samples were also found to be more detectable by the two indices than the 10% spuriously high aberrant response patterns. The detection rates of the 20% and the 10% spuriously high aberrant response patterns by the two indices were found to be higher under high item difficulty parameters, and were found to be low under the low item difficulty parameters. This is not surprising as it is expected that more responses are changed from incorrect to correct and fewer responses are changed from correct to incorrect under high item difficulty parameters. The 20% and the 10% spuriously low aberrant response patterns were also more detectable under the low item difficulty parameters because more responses are changed from correct to incorrect and fewer are changed from incorrect to correct under the low item difficulty parameters.

The detection rates of the two indices were also found to increase as a function of both ability distributions and item

difficulty parameters. Overall, LZ had higher detection rates of the 20% spuriously high aberrant response patterns than the corresponding detection rates by ECIZ4. Under combination of low ability distributions and high item difficulty parameters, ECIZ4 could detect 23%, 43%, 54%, and 71% whereas LZ could detect 26%, 47%, 57%, and 77% at the corresponding false alarm rates of 1%, 5%, 10% and 25% respectively. Under the same combinations LZ could detect 7%, 21%, 30% and 55% while ECIZ4 could detect 6%, 18%, 28% and 50% of the 10% spuriously high response patterns at the corresponding false alarm rates of 1%, 5%, 10% and 25% respectively. This suggessts that LZ index performs better than ECIZ4 index in detecting aberrant response patterns in the spuriously high treatment samples.

For the case of the 20% spuriously low aberrant response patterns, LZ had high detection rates of 33%, 54%, 65% and 83% while the corresponding detection rates of ECIZ4 were 36%, 56%, 64% and 78% under the combination of low item difficulty parameters and high ability distributions at the corresponding false alarm rates of 1%, 5%, 10% and 25% respectively. The corresponding detection rates of 10% spuriously low aberrant response patterns were 20%, 40%, 52% and 66% for LZ and they were 13%, 30%, 39% and 57% for ECIZ4.

The two indices had very low detection rates of spuriously low treatment samples under the combination of high item difficulty and low ability distributions. The detection rates of the 20% spuriously low aberrant response patterns by the two indices were

found to be similar. However, the detection rates of the 10% spuriously low aberrant response patterns by LZ were found to be higher than the corresponding detection rates by ECIZ4 for most of the combinations.

It is quite difficult to compare the detection rates obtained in this study with those obtained by other researchers because of a number of reasons. First, previous researchers used different experimental conditions from the ones considered in this study. A majority of them assumed that the examinee ability distributions are always normal (0, 1), a situation which is not always true. They also assumed that tests are always constructed to cover a wide range of item difficulty parameters. But different needs demand different tests. Hence, in this study, item difficulty and ability distributions were manipulated.

Secondly, most previous reseachers used different levels of aberrance. Noonan (1990) used 30%, 15% and 10%; Drasgow et al. (1985) used 10%, 20% and 30%. In this study only two levels of aberrance(10% and 20%) were considered. However, the high detection rates found in this study are consistent with the previous reported results. In the next chapter, the summary and the limitations of this study are presented.

CHAPTER V.

SUMMARY AND LIMITATIONS OF THE STUDY.

This study investigated two issues related to appropriateness measurement. The first issue concerned the effects of item difficulty and ability distributions on the distributional characteristics of LZ and ECIZ4 appropriateness indices in non-aberrant response patterns. The second issue concerned the effectivenes of the two indices under different combinations of item difficulty, ability distribution, type of aberrance, and level of aberrance. In this chapter the summary of the results and the limitations of this study are presented.

Summary of the Results.

The summary of the results of the distributional characteristics of the two indices in non-aberrant response patterns are as follows:

1. The mean, standard deviation, skewness, and kurtosis values of LZ and ECIZ4 were affected by the variations of item difficulty and ability distributions; they significantly differed among the nine combinations.

2. The mean, standard deviation, skewness, and kurtosis of LZ and ECIZ4 differed from the expected values.

3. The means and the standard deviations of LZ were closer to the expected values than the means and the standard deviations of ECIZ4 whereas the skewness and the kurtosis values of ECIZ4 were

closer to the expected values than the skewness and kurtosis values of LZ.

4. LZ had negative skewness values while ECIZ4 had positive skewness values. This suggests that LZ is skewed to the left and ECIZ4 is skewed to the right; i.e each index is skewed towards the direction of it's aberrance.

5. The distribution of LZ and ECIZ4 approximated a normal distribution when the ability estimates matched the item difficulty. The mean and the standard deviation values of the two indices were also closer to the expected values when the ability estimates matched the item difficulty parameters. The skewness and kurtosis values were also very low under these conditions. The opposite was true when the ability estimates did not match the item difficulty parameters.

6. The percentile estimates of LZ and ECIZ4 were also affected by the variations of item difficulty and ability distributions.

7. The percentile estimates of LZ were also closer to the expected values than the percentile estimates of ECIZ4.

8. Overall, LZ approximated a normal distribution better than ECIZ4.

The summary of the detection rates of LZ and ECIZ4 in aberrant response patterns are as follows:

1. The 20% spuriously low treatment samples were easier to detect by LZ and ECIZ4 than the 20% spuriously high treatment samples.

2. Spuriously low treatment samples were more detectable under

low item difficulty parameters and spuriously high treatment samples were more dectable under high item difficulty parameters.

3. The detection rates of LZ and ECIZ4 increased as a function of the level of aberrance. Twenty percent spurious treament samples were easier to detect than the 10% spurious treatment samples.

4. LZ had higher detection rates of spuriously high and spuriously low aberrant response patterns than ECIZ4 index under the high item difficutly parameters at low (0.01 and 0.05) false positive rates.

5. ECIZ4 had higher detection rates of the 20% spuriously low and the 20% spuriously high aberrant response patterns than LZ index and LZ had higher detection rates of the 10% spuriously low and the 10% spuriously high aberrant response patterns than ECIZ4 under the low and under the medium item difficulty parameters at low (0.01 and 0.05) false positive rates .

Recommendations to Practitioners.

Considering the results of the present study, the following recommendations can be made:

1. LZ could be used to detect spuriously low or spuriously high aberrant response patterns if a test consists of items with high item difficulties.

2. ECIZ4 index could be used to detect spuriously low and spuriously high aberrant response patterns if a test consists of items with low and moderate item difficulties.

3. Cutoff scores should be established using a large

population of examinees with a test which matches their ability. However, this study has shown that cutoff scores can vary according to different testing conditions. Therefore, test users should see to it that cutoff scores are reviewed regularly.

Limitations of the Study.

Simulated data were used in this study. Future researchers can replicate this study using real data. It was also assumed in this study that all the examinees reached and attempted all the questions; a situation which doesn't usually happen in real life. Future researchers can use data matrix containing omits.

Only one distribution of item difficulties (uniform) with varying intervals was considered. Future researchers could use skewed or normal distributions of item difficulties. Examinee ability were also restricted to normal distributions with different means but with the same standard deviation. However, it is possible to have other types of ability distributions in real life situations.

In addition, data were generated according to the three parameter model. One and two parameter models could also be used for future research. Spuriously high and spuriously low scores were analysed separately. In real life, a sample may have some examinees with spuriously high scores and others with spuriously low scores. This would presumably affect the detection rates.

Finally, the combined effects of test length, item difficulty and examinee ability on the distributional characteristics and the effectiveness of LZ and ECIZ4 should be investigated.

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.

Candell, L.G., & Levine, M.V. (1990). Detecting aberrant responses to the initial items on computerised Adaptive Tests. An application of appropriateness measurement. A paper presented at the annual meeting of the American Educational Research Association. Boston.

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

Donlon, T.F., & Fischer, F.E. (1968). An index of individual's agreement with grouped detrmined 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.

Dragow, F., & Guertler, E. (1987). A - decision - theoretic

approach to the use of appropriateness measurement for detecting invalid test and scale scores. Journal of Applied Psychology, 72(1), 10-18.

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

Drasgow, F., Levine, M.V., & McLaughlin, M.E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. Applied Psychological Measurement, Vol. 72, No.1, p.59-79.

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

Hambleton, R.K., & Rovinelli, R. A FORTRAN7 IV program for generating examinee response data from logistic test models. Behavioral Science, 1973, 18, 74.

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

Harnisch, D.L., & Tatsuoka, K.K. (1983). A comparison of

appropriateness indices based on item response theory. In R.K. Hambleton (Ed.), Appilcations of item response theory. In Vancouver, B.C.: Educational Research Institute of Brtish Columbia.

Hoijtink, H. (1986). Detecting aberrant response patterns in the unidimensional scaling model of Rasch. Unpublished manuscript. University of Groningen, Netherlands.

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

Jaeger, R.M. (1988). Use and effect of caution indices in detecting aberrant reponse patterns in standard setting judgements. Applied Measurement in Education, (J6), 17-31.

Kellett, R. (1989). An emperical investigation of Hui-Triandis model to assess the measurement equivalence of different Canadian subpopulation (Cultural groups). Unpublished dissertation, University of Ottawa.

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

Levine, M.V., & 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.

Noonan, B.W. (1990). The effects of test length, IRT model, type of aberrance, and level of aberrance on the distribution of three appropriateness indices. Unpublished dissertation , University of Ottawa.

Molenaar, I.W., & Hoijtink, H (1990). The many Null distributions of person fit indices. Psychometrika, vol. 55, 1, 75-106.

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

Tatsuoka, K.K (1984). Caution indices based on item response theory. Psychometrika, 49(1), 95-**219.**

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

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

Van der Flier, H. (1982). Deviant Response patterns and comparability of test scores: Journal of Cross-cultural Psychology, Vol.13, No.3, september 1982, 267-298.

Wood, R.L, Wingersky, M.S, & Lord, F.M. (1976). LOGIST - A computer program for estimating examinee ability and item characteristics curves (Research Memorandum No. 76 - 6). Pinceton, NJ: Educational testing.

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