EXAMINATION OF THE DISTRIBUTION OF THE LOGISTIC REGRESSION AND THE MANTEL - HAENSZEL STATISTICS UNDER DIFFERENT CONDITIONS OF THE NULL HYPOTHESIS: A MONTE CARLO STUDY

By Charles.M.O.Ochieng

Thesis presented to the School of Graduate
Studies and Research of the University of Ottawa
in partial fulfillment of the requirement for
the degree of Master of Arts in Education.

ACKNOWLEDGEMENTS.

I wish to thank Dr.Marvin Boss, my supervisor, whose guidance, suggestions,and encouragement provided me with a great deal of motivation to continue with my research.

I am grateful to Dr.Marc Gessaroli and Bruno Zumbo for sharing with me their abudant knowledge of statistics and data analysis. I owe a lot of gratitude to Len Fleming and Chris Currathers, senior computer consultant at the University of Ottawa, who modified the data generation programs to fit my requirements.

Finally, I wish to thank my wife and children for their support and patience.

ABSTRACT

Educators and practitioners have been striving for bias-free tests for the last few decades. As a result of this, several indices for bias detection have been developed, among which are the logistic regression and Mantel-Haenszel procedures. However, the effects of variables other than DIF on the performance of the logistic regression and Mantel-Haenszel indices have yet to be researched.

The present study examines the effects of sample size, item difficulty, item discrimination, and ability distribution on the distributions and percentiles (P_{90} and P_{95}) of logistic regression and Mantel-Haenszel statistics under the null hypothesis. Simulated data were used in order to evaluate the effects of the stated variables on the distributions of the logistic regression indices of uniform (LU) and nonuniform (LN) differential item functioning DIF or item bias. The same simulated data were used to evaluate the effects of the variables on Mantel-Haenzel procedure (MH-delta and MH-CHISQ).

Three sample sizes were used, each with three to one ratio (Reference/Focal) of 300/100, 600/200, and 1200/400. Three values of 'a' were simulated, 0.6, 1.0, and 1.4. A sixty three item test was simulated for each 'a' value with twenty one 'b' values of item difficulty ranging from -2.0 to +2.0 . For each 'b' value there were three items. Item response strings were simulated for focal and reference groups for equal and unequal ability distributions. Under the equal ability distribution both reference and focal group had a mean of zero and standard deviation of one. In the case of unequal ability distribution, the mean for the reference group was zero with a standard deviation of one, while in the focal group the

mean was -0.5 with a standard deviation of 0.83.

One hundred replications were generated for each combination of sample size, item discrimination parameter 'a', item difficulty 'b', and ability distribution. The twenty one values of 'b' were divided into three categories at equal intervals of low 'b', medium 'b', and high 'b'.

The four indices were computed and for each index the mean, standard deviation, skewness, kurtosis, P_{90}, and P_{95} were computed for each replication. In the case of MH-delta, absolute values were used in computing the two percentiles.

Taking sample size, item discrimination, item difficulty, and ability distribution as independent variables, a MANOVA was conducted on each of the four indices using the stated descriptive statistics as dependent variables, so as to analyse the effect of the independent variables on distribution of the indices. A MANOVA was conducted on P_{90} and P_{95} for each index, in order to determine the effects of the variables on the cutoffs at the two percentiles. Significant MANOVA was considered at $p < 0.05$. This was then followed by using ANOVA at $p < 0.01$ for all significant results of the MANOVAs.

Sample size showed a significant main effect for all the dependent variables except LU. Ability distribution showed no significant main effect for any of the four indices, indicating that it had no significant effects on the performance of any of the four indices.

Item discrimination significantly affected MH-CHISQ only for means.This variable significantly affected the distribution of MH-delta for the mean and standard deviation. The mean and standard deviations of MH-delta were greater at a high level of item discrimination than at a low level of item discrimination. P_{90} and P_{95} were also significantly

greater at a high item discrimination than at a low level of item discrimination.

In the case of the LU index item discrimination had no significant effects. However for LN index, item discrimination affected the distribution, P_{90}, and P_{95} .

Item difficulty showed significant effects at the mean and standard deviation but had no significant effect on the skewness and kurtosis of MH-CHISQ. Item difficulty also showed a significant effect at P_{90} but not at P_{95} of MH-CHISQ. In the case of MH-delta, item difficulty significantly affected each of the six dependent variables. For LU, item difficulty did not show any significant effects, while for LN item difficulty showed a significant main effect on the mean, standard deviation, P_{90}, and P_{95}.

At a low sample size and high discrimination level the P_{90} of MH-CHISQ was lower than the expected value of 2.70. The use of the computed P_{90} values would result in more false positives than when the expected value is used. However at any other level of the independent variables, the P_{90} values of MH-CHISQ were over estimated and were larger than the expected value.

For MH-delta, the means were close to the expected value of zero. For the LU index the means were close to the expected value of one while the standard deviations were slightly lower than the expected values. The P_{90} and P_{95} obtained for LU were larger than expected values of 2.70 and 3.84 respectively. As for the LN index the means and standard deviation were close to the expected values, although at high discrimination levels and high item difficulty the standard deviations were overestimated.The P_{90} and P_{95} values were greater than the tabled values at all the levels of independent variables for LN.

Results of this study show that the logistic regression procedure has advantages over

the MH procedures, taking into account the effects of the independent variables studied. This is evident from the fact that the distribution of LN and LU index are known and that the four independent variables had no significant effect on the LU index.However the observed values were notably larger than expected values. Further research should be done to evaluate the effects of the stated variables and others such as test length, and using data with known amount of dif. Generalization of this study should be proved by replications of its findings. Evidently, variables other than DIF, significantly influence the two procedures.

TABLE OF CONTENTS

Pages

LIST OF TABLES viii

CHAPTER I:INTRODUCTION ..... 1.

The Mantel-Haenszel procedure ..... 2.

The MH-chisquare index (MH-CHISQ) ..... 5.

The Mantel-Haenszel alpha (MH-alpha) ..... 6.

The Mantel-Haenszel delta (MH-delta) ..... 7.

The Logistic Regression(LR)procedure ..... 9.

CHAPTER II: LITERATURE REVIEW ..... 12.

Summary of the Findings ..... 36.

Purpose of the Study ..... 39.

Research Questions ..... 41.

CHAPTER III: METHODOLOGY ..... 42.

Data Collection Approach ..... 42.

Assumptions Made in the Study ..... 44.

Characteristics of the Items ..... 44.

Characteristics of the Examinees ..... 45.

Simulation Model and Data Analysis Procedure ..... 46.

Analysis of the Effects of the Variables on the Distribution of the Indices...46.

Analysis of False Positives.....47.

CHAPTER IV: RESULTS AND DISCUSSION.....48.

Effect of the Independent Variables on the Distribution of MH-CHISQ.....49.

Effect of the Independent Variables on the P_{90} and P_{95} of MH-CHISQ.....54.

Effect of the Independent Variables on the Distribution of MH-delta.....60.

Effect of the Independent Variables on the P_{90} and P_{95} of MH-delta.....66.

Effect of the Independent Variables on the Distribution of LU .....70.

Effect of the Independent Variables on the P_{90} and P_{95} of LU.....72.

Effect of the Independent Variables on the Distribution of LN.....73.

Effect of the Independent Variables on the P_{90} and P_{95} of LN.....78.

CHAPTER V: SUMMARY AND CONCLUSIONS.....83.

Relationship of Findings to other Research.....86.

Limitations and Further Research.....89.

Educational Implications.....90.

REFERENCES:.....93.

LIST OF TABLES Pages

Table 1: Frequencies of Responses of Focal and Reference groups at jth Ability level....4.

Table 2: Conditions under which the four Indices were Studied Using a 63 Item Test for each of the two Ability Distributions over 100 Replications....43.

Table 3: Group Means of the MH-CHISQ Means for the Discrimination by Item Difficulty two-way Interaction....50.

Table 4: Group Means of the Standard Deviation of MH-CHISQ for the Discrimination by Ability Distribution Interaction....52.

Table 5: Group Means of P_{90} of MH-CHISQ for the Interaction of Ability Distribution by Item Difficulty....55.

Table 6: Group Means of P_{95} of MH-CHISQ for the two-way Interaction of Discrimination by Ability Distribution....57.

Table 7: Descriptive Statistics for MH-CHISQ over 100 Replications across the Independent Variables....59.

Table 8: Descriptive Statistics of MH-delta over 100 Replications across the Independent Variables....61.

Table 9: Group Means of Skewness of MH-delta for the Discrimination by Ability Distribution Interaction....65.

Table 10: Group Means of the MH-delta P_{90} for the three-way Interaction of Sample Size by Item Difficulty by Discrimination.....67.

Table 11: Group Means of P_{95} of MH-delta for the four-way Interaction of Sample Size by Discrimination by Item Difficulty by Ability Distribution.....68.

Table 12: Descriptive Statistics of LU across the Independent Variables over 100 Replications.....71.

Table 13: Descriptives Statistics of LN across the Independent Variables over 100 Replications.....74.

Table 14: Group Means of LN for the Discrimination by Item Difficulty Interaction.....75.

Table 15: Group Means of the Standard Deviation of LN for the two-way Interaction of Item Difficulty by Discrimination.....77.

Table 16: Group Means of P_{90} of LN for the Discrimination by Item Difficulty two-way Interaction.....79.

CHAPTER I

INTRODUCTION

As a result of social, legal, and political concerns expressed over the last two decades, test developers have been striving to develop suitable procedures for detecting and correcting any bias that may exist in tests. This has been a major issue in educational and psychological measurement.

There are two major forms of bias in testing. Bias may occur at the test level. This is known as test bias, referred to as the use of an instrument as a criterion of decision making when the performance of one group is not predicted as accurately as the performance of another. Cleary (1968) stated that a test is biased for members of a subgroup of a population if consistent nonzero errors of prediction occur for its members. In other words bias can be defined as the unfair use of a test to predict future performance of a subgroup or subpopulation.

The other major form of bias is item bias . At the item level, item bias is said to occur, if for members of two groups of equal ability the probability of obtaining a correct response for a given item is not the same for each group(Crocker & Algina,1986).

Two notions central to the concept of item bias are that examinees performance on an item may be influenced by variables other than differences in ability on the dimension of interest. The second one is that these variables may affect the performance in a way that differs systematically for some subpopulation which gives an unfair advantage to one group over another (Crocker & Algina, 1986).

The occurrence of item bias will influence the precision of measurement and interpretation of the test when administered to different groups of a population. The term item bias has been replaced in some bias studies by the term, differential item functioning, (dif) (Holland & Thayer, 1986), to describe the phenomenon of different subgroup's reaction to certain items in a given test. The two terms, item bias and dif, may be used interchangeably.

There are judgemental and statistical procedures that have been developed and used to detect dif. However, statistical methods have been found to be less subjective and more reliable.

Several indices and procedures have been developed, among which are the Mantel-Haenszel procedure (Holland & Thayer, 1986) and more recently, the logistic regression procedure (Swaminathan & Rogers, 1990). Due to the popularity and common use of the Mantel-Haenszel procedure as well as the promising outlook for logistic regression in dif studies, there is a need to study the indices to determine if they are affected by variables other than dif. Each of these indices will now be discussed.

The Mantel-Haenszel Procedure

Originally the Mantel-Haenszel procedure was developed for studies in retrospective epidemiological investigations of the relationship between the presence or absence of a potential risk factor and the occurrence of disease (Mantel & Haenszel, 1959). However, this procedure was adopted by Holland (1985) for the detection of dif thus providing an alternative to IRT (item response theory) methods that are expensive

and require strict assumptions (Holland & Thayer, 1986; McPeeck & Wild, 1986). Since then, as a result of its computational simplicity and relatively low cost, the Mantel-Haenszel procedure has gained popularity and is now widely used to study dif. Various examination boards have adopted this procedure, e.g. ETS (Educational Testing Services, Princeton, New Jersey), to study and detect dif.

The Mantel-Haenszel indices (MH-indices) fall into three main categories; the MH-chi square, the MH-alpha (also known as common odds ratio), and the MH-delta. Of these indices, the MH-delta is more popular and appealing as a result of its practical and interpretational advantage, as will be seen later. In the MH-procedures, the performance of the focal group is of primary interest while that of the reference group is considered the standard against which the comparison of the focal group's performance is made (Holland & Thayer, 1986).

For each ability level, the MH-procedure has a contingency table of the form shown in Table 1. In the table, A_j denotes the number of members in the reference group R with the correct response to the studied item at ability level j, while B_j denotes the number of members in the reference group R, with incorrect response to the item. C_j denotes the number of members in the focal group with the correct response to the studied item at ability level j and D_j is the number of members in the focal group with an incorrect response to the studied item at ability level j.

Table 1

Frequencies of Responses of Focal and Reference groups at a given Ability level j (Holland & Thayer, 1986)

	Response on the studied item		
Groups	Correct(1)	Incorrect(0)	Total
Reference Group R	A_j	B_j	n_{Rj}
Focal Group F	C_j	D_j	n_{Fj}
Total	M_{1j}	M_{0j}	T_j

M_{1j} is the number of examinees in the examinee population with the correct response to the studied item at ability level j, while Mo_j is the number of examinees with incorrect response to the studied item at ability level j. The total number of examinees at ability level j is denoted by T_j, while n_{Rj} is the total of the reference group members and n_{Fj} that of the focal group, who responded to the studied item at ability level j. Each of the three MH-indices can be derived from the 2 by 2 contingency tables for each ability level, as shown in Table 1 (Holland & Thayer, 1986).

The MH-chi square index(MH-CHISQ)

The MH-chi square statistic is computed using the cell frequencies in the 2 by 2 table shown in Table I, as follows:

$$\text{MH-CHISQ} = \frac{\left(\left| \sum_{j=1}^{k} (A_j) - \sum_{j=1}^{k} E(A_j) \right| - 1/2 \right)^2}{\sum_{j=1}^{k} \text{var}(A_j)} \qquad \text{Equation 1}$$

where $E(A_j) = n_{Rj} . M_{1j}/T_j$

and $\text{Var}(A_j) = \frac{n_{Rj} . n_{Fj} . M_{1j} . M_{0j}}{T^2_j (T_j - 1)}$ where j is the ability level

In dif detection studies the matching criterion upon which comparable members of the focal and reference groups is based on the ability level. It is assumed that comparable members of the two groups are of equal ability. Once the criteria for matching has been selected, the data for the studied items for the examinees in the reference group and focal group may be arranged into a series of k 2 by 2 tables similar to Table 1, where k is the number of ability scores.

In their study, Holland and Thayer (1986) suggest that, in order to state a statistical hypothesis based on the MH procedure, it is necessary to have a sampling model for the data given in Table 1, where it is assumed that the marginal totals are fixed and the data for reference group and focal group are random samples of sizes n_{Rj} and n_{Fj} from matched subpopulations of the reference and focal groups. It then follows that A_j and C_j are independent binomial variates with parameters (n_{Rj}, P_{Rj}) and (n_{Fj}, P_{Fj}) respectively where P_{Rj} is the probability of correct response on the studied item at ability

level j of the reference group members, and P_{Fj} is the probability of a correct response for members of the focal group, at the jth ability level.

The values of q_{Rj} and q_{Fj} are the reference and focal groups respective probabilities of incorrect response to the studied item at the ability level j.

The hypothesis of no dif corresponds to the null hypothesis;

$H_0 : P_{Rj} = P_{Fj}$ Where $_j$ is the ability level Equation 2

The MH-CHISQ, under the null hypothesis H_0, has an approximate chi-square distribution with one degree of freedom. The hypothesis, H_0, is also the hypothesis of conditional independence of group membership and the score on the studied item given the matching variable as indicated in Bishop, Fienberg, and Holland (Holland & Thayer 1986). Total test score is frequently used to estimate ability which is the matching variable in the Mantel-Haenszel procedure.

The Mantel-Haenszel alpha (MH-alpha)

The MH-alpha tests for the equality of the ratio of probabilities of success (correct response) to probabilities of incorrect response for each item by the focal and reference group. This is deduced from equation 3 as shown.

$$\frac{P_{Rj}}{q_{Rj}} = \alpha \frac{P_{Fj}}{q_{Fj}}$$ where $_j$ is the ability level. Equation 3

When α is equal to 1, there is no dif. This corresponds to the null hypothesis. The alternative hypothesis is, when α is not equal to 1.This leads to rejecting the null hypothesis indicating that there is dif.

The value of MH-alpha (common odds ratio, α) is the average by which the odds that a member of the reference group is correct on the studied item at ability level j

differs from the corresponding odds for a comparable member of the focal group. Values of MH-alpha that exceed one, correspond to items on which the reference group performed better on average than did comparable members of the focal group. When MH-alpha equals one, this corresponds to items for which the focal group performed as well as the comparable members of the reference group.Whenever alpha is less than one, this corresponds to items on which the focal group performed better than the reference group.

The MH-alpha can also be derived from the contingency table (see Table I) so that:

$$\text{MH-alpha} = \alpha = \frac{\sum_{j=1}^{k} A_j D_j / T_j}{\sum_{j=1}^{k} B_j C_j / T_j} \qquad j = 1,2,3...k \qquad \text{Equation 4}$$

The Mantel-Haenszel delta (MH-delta)

The MH-delta is a transformation of the common odds ratio. The MH-delta values correspond to the difference in difficulty of a studied item with respect to two groups, the reference and focal groups. This index is the difference in item difficulty for the studied item over all ability levels for the two groups. The transformation of common odds ratio to MH-delta is done by taking the logarithm of the common odds ratio and then multiplying it by the value,-4/1.7 so that:

$$\text{MH-delta} = -4/1.7 \ \ln(\text{MH-alpha}) = -2.35 \ln(\text{MH-alpha}) \qquad \text{Equation 5}$$

The MH-delta has the interpretation of being a measure of dif in the scale of differences in item difficulty as measured in the ETS difficulty scale (Educational Testing Services;

Princeton, New Jersey). The index upon which the MH-Delta is derived is the MH-Z which is a logarithmic transformation of the MH-alpha so that:

$$\text{MH-Z} = \frac{-1}{1.7} \ln (\text{MH-alpha}) \qquad \text{Equation 6}$$

The MH-Z and the MH-delta are similar in all practical aspects. While the delta difficulty scale has a standard deviation of 4 and a mean of 13, the Z value has a mean of zero and a standard deviation of one. However, the MH-Z and MH-delta have an expected value of zero and unknown standard deviation. Both MH-Z and MH-delta are equal to zero if no dif is presesnt. Their distribution is unknown.

The MH-indices allow for careful matching of examinees on a relevant criterion besides providing a single summary measure of the magnitude and direction of the dif exhibited by the studied item. The MH-Z and the MH-delta indices are useful in that they have the advantage of indicating which group is disadvantaged by the item studied.

MH-CHISQ is a general extension of the chi-square statistical test. The fact that it provides a single degree of freedom chi-square test, means that it is a very powerful test against the alternative to the null hypotheses. The second procedure to be discussed is logistic regression (LR), a description of which is now given.

The Logistic Regression (LR) procedure

This is a model based approach to dif detection. The regression model is based on a probability function used to predict item response from group membership and ability of examinees. The dif statistics in LR are based on testing regression coefficients for statistical significance. The probability function (upon which the model is based) used to predict item response is as follows:

$$P(u = 1|\Theta) = \frac{\exp(\beta_0 + \beta_1\Theta)}{1 + \exp(\beta_0 + \beta_1\Theta)} \qquad \text{Equation 7}$$

Where u = item score which takes value 1 for a correct response and 0 for an incorrect response.

and Θ = ability of the examinee observed,

β_o = the intercept parameter,

β_1 = the slope parameter.

Equation (7) above is the standard logistic regression (LR) model for predicting a dichotomous dependent variable from a given independent variable, (Swaminathan & Rogers, 1990).

The model given in equation (7) can be used in detecting dif by specifying two equations for the two groups of interest, namely the reference and focal groups. This can be expressed as:

$$P(u_{ij} = 1|\Theta_{ij}) = \frac{\exp(\beta_{0j} + \beta_{1j}\Theta_{ij})}{1 + \exp(\beta_{0j} + \beta_{1j}\Theta_{ij})} \qquad \text{Equation 8}$$

Where $i = 1, 2, 3 \ldots n_j$
and $j = 1$ or 2

Here u_{ij} is the response of person i in group j to the studied item. β_{0j} for group j, is the intercept parameter while β_{1j} is the slope parameter for group j. The term, Θ_{ij} is the ability of examinee i in group j. P is the probability of the response u_{ij} (Hambleton & Swaminathan, 1985).

From this model dif is said to occur if, for a studied item, examinees of the same ability but from different groups do not have the same probability of correct response on the item (Hambleton & Swaminathan, 1985). It follows that no dif is present when the logistic regression curves for the two groups are the same, i.e. the slope and intercept are the same for the two groups.

Uniform dif is said to occur when one group consistently performs better than the other group over all ability levels, for the studied item for matched ability levels. In other words the slope parameter or parameter 'a' is equal for the two groups while the item difficulty parameter is not equal for the two groups. Nonuniform dif occurs when the slope or discrimination parameters for the two groups are not equal over all ability levels while the item difficulty parameters may or may not be equal for the studied item. This means that one group may not consistently perform better than the other in the population, across all the ability levels (Mellenberg, 1982).

From equation (8), whenever $\beta_{11} = \beta_{12}$ (equal slopes) and β_{01} is not equal to β_{02}, the logistic regression curves are parallel and so uniform dif is said to occur. Whenever $\beta_{01} = \beta_{02}$ but β_{11} is not equal to β_{12}, the curves are not parallel and hence the presence of nonuniform dif can be deduced. In nonuniform dif there is interaction between group

membership and ability for a studied item. So as to carry out tests of hypothesis of interest for uniform and nonuniform dif, a statistical test of significance based on the distribution theory of the model is considered. The chi-square test of significance with two degrees of freedom is conducted for the logistic regression statistic for dif. However, one degree of freedom chi-square tests can be conducted for uniform dif and nonuniform dif independently.

CHAPTER II

LITERATURE REVIEW

Educators and test developers have been striving to develop bias free tests for the two decades.Several indices have been proposed, among which are the MH procedures and the LR procedures. These two procedures have attracted the attention of the test developers because the indices have shown promise,they are cost effective and do not require large sample sizes as well as strict assumptions, unlike IRT (item response theory) methods. Added to this, is the fact that most of the current research on dif indices has been on signed indices which may not be effective in detecting nonuniform bias. Nonuniform bias is well detected by LR and to a lesser extent by MH procedure. The LR procedure and the MH procedure provide omnibus tests for dif that is both difficulty sensitive and differential sensitive. Several studies to investigate various factors that influence LR and MH statistics and resulting distributions have been conducted using real and simulated data.These studies are suitably conducted if the test data have known dif conditions. This is not common for most real data and so simulated tests lend themselves to these investigations.

Numerous studies have been conducted to assess the characteristics of the MH-statistics using data from real tests. However, some difficulties arise in their use since the true state of dif is not known. This is a limiting factor in that it is not possible to assess with any certainty the power and distribution of the MH and LR indices. For this reason simulated tests have been used to study factors that influence dif statistics and their distributions.

Mazor, Clauser, and Hambleton (1991) conducted a study based on simulated data, so as to examine the effects of sample size on the detection rate of the MH-statistic under different values of the discrimination parameter, 'a',and item difficulty parameter, 'b'. This was done for equal and unequal ability distributions of the reference and focal groups.

Five different 75 item tests were generated for each group by first generating 59 items that were non-biased and common among the five tests. Eighty additional items, different from the 59 item tests, were generated in a set of five subtests consisting of 16 items with different levels of discrimination and item difficulty under biased conditions. Of the 16 items per subtest, four values of 'a' were used (0.25, 0.6, 0.9, and 1.25) .To simulate dif, five values of 'b' were used and set to differ by 0.25, 0.5, 1.0 and 1.5 for reference and focal groups.Thus, five levels of item difficulty were crossed with four levels of discrimination and four differences in item difficulty. This resulted in eighty dif items. The eighty items were divided into five tests of 16 items each. The five sets of 16 items were then combined with five sets of 59 item tests resulting in five 75 item tests. The 'c','a', and 'b' item parameters of the 59 items common to each test were based on a previous GMAT test adminstration. The 'c' parameter was set at 0.2 for all items.

Since the underlying ability was hypothesised to affect the performance of MH-statistics, three samples of 2000 examinees each were generated such that the first two sets had the same ability distribution with a mean of zero and a standard deviation of one; one was used as the reference group,the other as focal group(focal group 1). The third set was generated such that the ability distribution of the reference group had

a mean of zero and a standard deviation of one while that of the focal group had a mean of -1.0 and a standard deviation of one. The third (focal group 2) set was used to compare groups of unequal ability distribution while the first and second samples were used to compare groups of equal ability distribution.

For each of the three groups, DATAGEN (Hambleton & Rovenelli, 1973) was run for 2000 examinees, and a sample of 1000 was randomly selected. The process was repeated for 500, 200, and 100 examinees. The MANTEL computer program was run so as to compute MH-statistics at the stated sample sizes. For the 1000 and 500 sample sizes one replication was done for each of the equal and unequal ability distribution conditions. For 200 and 100 sample size runs, each was replicated twice for each of the ability distribution conditions.

Results reported were based on a level of significance of 0.01. Items identified as having dif on the first run were removed for the calculation of the overall test score and then the MH-statistic was recalculated for each of the tests, as suggested by Holland and Thayer (1986).

The percentage of dif items correctly flagged decreased significantly as the sample size decreased. At the 2000 sample size, 64 percent of the dif items were correctly identified for unequal ability distribution samples. For the same sample of 2000 at equal ability distribution, 74 percent of dif items were correctly identified. At the 1000 sample size and unequal ability distributions ,58 percent of the dif items were correctly identified, while at equal ability distributions for the same sample size 61 percent of dif items were correctly identified. In general, when the ability distributions for reference

and focal groups were equal, correct identification rate was higher than when the ability distributions were unequal. This trend was observed to be even more so, with smaller sample sizes.The correct identification rate also diminished with smaller sample sizes.

Poorly discriminating items were least likely to be correctly identified except for large sample sizes and at greater differences on item difficulty for the two groups. Items with high item difficulty were least likely to be correctly identified compared to items of moderate item difficulty and high discrimination. Items of moderate item difficulty and high discrimination were easily identified. Comparing groups with unequal ability distributions resulted in items of large p-values differences being identified across all the sample sizes studied. With equal ability distribution more items of small differences were detected although the rate decreased with the decrease in sample size.

It has been hypothesized (Donoghue & Allen, 1991) that various forms of pooling or categorisation of score groups as matching variables on the MH-procedure has an effect on the MH statistic and its detection rate of dif items. Donoghue and Allen (1991) conducted a Monte Carlo study to investigate the effects of various types of pooling or forming of the matching variable on the MH-procedure. Two main forms of pooling or matching were studied, namely, 'Thick' matching and 'Thin' matching.

Independent variables manipulated in the study were; test length, sample size, methods of forming matching variables, dif condition in the test items, discrimination parameter 'a', and item difficulty parameter 'b'.

The study design consisted of four levels of the test length (5, 10, 20, and 40 items), three levels of sample size (400, 800, and 1600) such that one quarter of each

sample size was from the focal group.There were three levels of 'a' (0.3, 1.0, and 1.5) and seven levels of 'b', item difficulty parameters, (-1.5, -1.0,-0.5, 0.0, 0.5, 1.0 and 1.5). So as to induce dif, item difficulty parameters were set to differ for reference and focal groups by adding 0.3 to those of the focal group, resulting in item difficulty in favour of the reference group. The 'c' parameter was set at 0.2 for all items. The distribution of the parameters was selected and simulated to approximate the empirical SAT (Scholastic Aptitude Test) item parameters given in Lord's study (Donoghue & Allen, 1991).

The three levels of sample size and four levels of test length were crossed to define 12 test conditions. Simulated item responses were then generated, according to a three parameter logistic model, and MH-analysis was performed using a FORTRAN 77 program (developed by the authors). Twenty replications for each of the 12 test conditions were made.

In "thin" matching variable, levels of 2 x 2 tables were based on all possible total test scores. "Thin" matching was done in each of the four test lengths. "Thick" matching was divided into several categories. This was done by forming a matching variable by pooling total score levels. The categories were; equal interval, percent of total sample, percent of focal sample, censored matching, minimum cell frequency, and the 'no' matching category.

The three MH-indices computed were MH-delta, the standard error of MH-delta, and MH-CHISQ. The mean and standard deviation of each of these indices for each of the conditions were used as the dependent variables in the study. This assisted in analyzing the effects of various manipulations on the dif measures obtained.

Results indicate that 'thick' matching yielded better results than 'thin' matching with small sample sizes. This was also true with short tests. With large sample sizes, 'thin' matching was better than 'thick' matching. This was also noted to be true for longer tests when the MH-delta index is used. In the case of MH-CHISQ, increase in test length and in sample size had an effect on the index. At equal intervals, correct identification of dif items was as good for 'thick' matching as for 'thin' matching. 'Thick' matching based on an approximately equal number of examinees per score interval and an equal number of focal group members per score interval yielded better results. There were no meaningful differences on the MH-indices between 'thin' and 'thick' matching with tests of moderate length. The results also indicated that parameter 'a' and parameter 'b' had significant effects on the indices depending on the matching variable chosen and the values of the parameters considered.

Gutierrez (1989) conducted a Monte Carlo study to examine the effects of sample size, discrimination parameter 'a', and item difficulty parameter 'b' on the distribution of the MH-Z under the null hypothesis. The study also examined the effects of these variables on the cutoffs established at P_{95}, $P_{97.5}$, P_5, $P_{2.5}$ at the distribution of the index which represented 0.05 and 0.01 false positive rates.

Three sample sizes (600, 1200, and 1800) were selected for study. In each sample size, 75 percent were members of the reference group while the rest were from the focal group. Three values of the discrimination parameters 'a' (0.7, 1.0, 1.3) were simulated. In each case the value of 'a' was common to all items of a test. Item difficulty parameter 'b' values ranged form -2.0 to +2.0, taken at an interval of 0.1. Since it was assumed

that no guessing took place in the simulated test, the 'c' parameter for guessing was set at zero and was not considered in the study. The simulated test consisted of 40 items.

One hundred replications were generated for each combination of sample size and 'a' parameter value. The mean and standard deviation of the MH-Z values were computed across the 40 items for each of the hundred replications. Four cutoff points at the mentioned percentiles were established.

A 3 by 3 ANOVA (sample size by 'a' value) was conducted over 100 replications on the standard deviation of the MH-Z index computed, for the 40 items. A two way interaction between sample size and value of 'a' was significant at 0.05 level. A simple effects analysis was conducted for each of the independent variables and Scheffe post hoc comparison done to determine which levels of the independent variables differed significantly.Certain combinations of 'a' values and sample sizes were found to have a significant effect on the standard deviation of the MH-Z. As sample size increased, the standard devation decreased, and as the value of 'a' increased, so did the standard deviation.

A repeated measures ANOVA was also conducted on the standard deviations of the MH-Z across the 100 replications of each combination of conditions.There was a significant interaction between sample size and 'a'value for the standard deviation of the MH-Z.

An analysis of the mean of the standard deviation of MH-Z for each of three levels of 'b' was conducted. The three way interaction between sample size, value of 'a', and value of 'b' was shown to be significant at 0.05 level. In all combinations of sample

size and value of 'a', the standard deviations were significantly larger when the 'b' values (item difficulty) were located at the extremes of the 'b' distribution than when 'b's were at the centre.

A MANOVA was carried out on the four cutoff percentiles. Sample size, discrimination, and the interaction of sample size and discrimination were found to be significant at 0.05 level. A simple effects analysis was then conducted. The three way interaction between sample size, value of 'a' and value of 'b' was shown to be significant at 0.05 level. In all combinations of sample size and value of 'a', the standard deviations were significantly larger when the 'b' values (item difficulties) were located at the extremes of the 'b' distribution than when 'b's were at the centre.The findings show that MH-Z is unstable at the extreme levels of item difficulty.

Computer based simulation studies have numerous advantages in the investigation of characteristics and distribution of dif indices, one of which is the fact that variables suspected to influence the indices can be manipulated. The dif studies can also be conducted under controlled conditions. Rogers and Hambleton (1989) carried out a study to evaluate computer simulated baseline statistics used in item bias studies.

Real and simulated test data were compared, with the real data serving as a validation to the simulated data. In the study (Rogers & Hambleton, 1989) simulated sampling distributions of three dif statistics were generated under the null hypotheses of no dif. This was done to determine cutoff points for use in baseline studies.The three dif statistics studied,were;IRT area method as suggested by Rudner,Getson and Knight (cited in Rogers & Hambleton, 1989), the root mean square difference method, and the

MH-alpha index.

Item parameter and ability parameter estimates were obtained from a test administered to 937 examinees (451 males and 486 females) who sat for the Cleveland Reading Competency Test. Simulated item responses were based on a three parameter logistic model. The 'c' parameter was set at 0.2. The objective was to simulate examinee item score data that reflected the actual examinee item data. Any differences in ability between males (reference) and females (focal) group were retained because ability estimates obtained from the analysis of real data were used in the simulation.Data were then generated.

For comparison of reference and focal groups,and the analysis of the data, real data of the examinees were split into four subgroups of two male groups (halved), M1, and M2, and two female groups,F1 ,and F2, based on the abilities of the examinees.These four groups were randomly equivalent. Similarly, the simulated data were also divided into four subgroups of two male groups ,M1, M2, and two female groups, F1, F2.

In the real data, comparisons of the groups were conducted as follows;M1 with M2, F1 with F2, M with F, M1 with F1, M2 with F2, (the M2 with F2 comparison served as a replication of the comparison of M1 with F1 samples). In the five sets of comparisons, the three dif statistics were computed (without dif). The distribution of these statistics served as a basis for evaluating the distribution generated for the simulated data.

For the simulated data the following comparisons were conducted;M1 with F1,

M2 with F2 and M with F. These distributions of dif statistics were compared to the matched set of the real data. The analysis of all the comparisons involved setting cutoff scores with real and simulated sampling distribution of the three dif statistics under the null hypothesis.This was followed by comparing the effects of the different cutoff scores on the number of items labelled as potentially biased.

Two analyses were conducted. The first one was conducted to evaluate the advantages of the selected simulation method. The second one was done to evaluate the advantages of utilizing simulated rather than real data. Sampling distribution in establishing cutoff scores and combined distributions from simulated data were compared to those of real data (M with F for both simulated and real data), in an attempt to assess how viable computer generated sampling distributions were.

The goodness of fit results indicated a close fit between the three parameter logistic models for simulated and real data. The curves plotted for each comparison of real and simulated data indicated no meaningful differences in simulated and real data results. This implied that simulated data closely approximated results from real data. These outcomes provide a sound rationale for using simulated data in studying characteristics and distributions of dif indices and in particular the MH-statistics.

Studies have been conducted to compare MH-statistics with other dif indices and methods of detecting dif. Camilli and Smith (1990) conducted a study to compare the accuracy and power of MH-CHISQ with the Randomisation test and the Jackknife test procedures of detecting dif in test items. The two procedures (Randomized and Jackknifing) make weaker distributional assumptions under the conditions of small

sample sizes and differences in ability distributions of reference and focal groups.

Subjects consisted of 1385 examinees (1085 whites, 300 blacks) who sat for a 30 item mathematics computation test based on the New Jersey Basic Skills Placement Examination (1984 version). Based on the item parameters of the real test data, two independent tests were simulated. A three parameter logistic model, fitting the real data, was used to generate item responses for the tests. The 'c' parameter estimate was conducted based on choices of options in the items that were least chosen by examinees with total scores of 10 or less. The 'a' parameters and 'b' parameters were obtained by use of a joint maximum likelihood algorithm developed by the authors. Ability estimates of the examinees were obtained using IRT. The means and standard deviations of the reference group and focal group raw scores and IRT estimated abilities were then computed.

Bias was induced in the simulated tests by adding specified values to the 'b' parameters of the focal group so as to fit the real data which had biased items. The MH-CHISQ statistic was computed for both simulated tests.

The Randomization test procedure based on Fisher's work (Camilli & Smith, 1990) was conducted for the two tests. Percentiles based on the Randomization test (randomised percentiles) were then obtained for each test. For the MH-CHISQ statistics computed for the simulated tests, nominal percentiles were also obtained. The two percentile procedures namely, the randomised percentiles and nominal percentiles of the MH-CHISQ were compared for the two simulated tests. A similar comparison was done for the real data test.

An additional analysis was done to compare the sensitivity of the randomized statistic to that of MH-CHISQ statistic at lower levels of dif. To simulate these conditions, bias differences ('b' values) added or subtracted when inducing bias in the first pair of simulated data were progressively reduced in three stages, by a half, a quarter, and finally to 'no' difference or zero and each time new data were generated using the same ability parameters and analyzed. The zero difference or no bias condition provided a baseline for interpreting increments in the indices. Once again the percentiles of the two indices were compared in each of the three conditions. The results suggested that MH-CHISQ statistic was moderately sensitive to small amounts of dif in small sample sizes. MH-CHISQ statistic was even more sensitive to highly discriminating items.

For the comparison of the MH-CHISQ statistic to the Jackknife test statistic only real data were used.The Jackknife procedure used was based on the Mosteller and Tukey Study (Camilli & Smith 1990).Significant tests were conducted on the item difficulties for the reference and the focal group and the resulting t-statistics were computed into percentiles. These percentiles were then compared to the MH-CHISQ nominal percentiles. The correspondence was close enough to suggest that the Jackknife test gives very similar results to the MH-CHISQ. It was concluded that no meaningful differences existed between the results for the indices.

There have been previous studies comparing MH-statistics to other indices.Spray (1989) compared the performance of three conditional dif statistic namely the MH statistic (MH-alpha), the STD statistic (The Standardised difference in proportion

correct) and the RMWSD (Root Mean Weighted Squared Difference in proportion correct), under simulated test conditions. The comparison of the indices was done in relation to their corresponding asymptotic dif indices.

Since the dif statistics were conditioned on total test score, the asymptotic indices based on these statistics were conditioned on latent ability. This was assumed to be continuous rather than discrete, ranging from minus infinity to positive infinity.

Three tests previously administered in the ACT assessment program were used in the study to compute the asymptotic indices. The three tests were; Mathematics Usage Test A of 40 items, Social Studies test B of 52 items, and another Mathematics Usage Test C of 40 items.

For Test A, 4000 examinees(2000 whites,2000 blacks) were drawn from a national population . For test B a similar sampling procedure was done. In both tests whites formed the reference group while blacks formed the focal group. For test C a sample of 4000 examinees (2000 males, 2000 females) was sampled randomly using a stratified sampling plan from the national population.Males formed the reference group while females formed the focal group.

So as to compute the asymptotic dif indices the item parameters from the real data (test A,B, and C) were re-estimated. The parameter estimates obtained for the three calibrations of each focal group were rescaled to their corresponding reference group using linear transformations. The ability distributions for each test and group were re-defined using ability density functions such that for test A, ability was normally distributed with a mean of -0.5 and a standard deviation of 1.0 for the focal group, while

for the reference group the mean was zero and standard deviation was one. For test B the same ability distributions were used.In test C the distributions of the focal and reference groups were the same, with a mean of zero and a standard deviation of one. A combined density function of ability distribution was used in the evaluation of the asymptotic MH-alpha index.

Two levels of sample size ratios (ratio of focal to reference samples) were used in the study, namely the ratio 1:1 and the ratio 1:10. Six conditions of the 1:1 ratio were used to analyze the asymptotic indices for each test. the conditions were : 2000/2000, 1000/1000, 500/500, 250/250, 100/100 and 50/50 for focal/reference groups.

For the ratio of 1:10, three conditions were analyzed for each test. The three sample ratio conditions were: 200/2000, 100/1000, 50/500 for focal/reference group. Asymptotic indices of the three dif statistics (asymptotic STD, asymptotic RMWSD, asymptotic MH-alpha) were computed and analyzed for all the nine sample size conditions for each test.The detection of dif items by the asymptotic indices was determined by selected cutoff values for each of the three asymptotic indices. It was noted that varying the ratio of focal to reference group samples affected only the MH-alpha asymptotic index and not the other two indices. The RMWSD asymptotic index identified both nonuniform and uniform dif. Nonuniform dif was not identified by the other indices because the MH-alpha asymptotic index and STD asymptotic index are directional and could only detect uniform dif.

For simulated tests A, B, and C, the item parameters were derived from the real tests A, B, and C. The ability estimates were obtained by randomly sampling ability

values from ability distributions of the focal groups and those of the reference groups based on the ratio of sample size of focal to reference groups. Item parameters were generated using a three parameter model. The generated data were used to compute the MH-alpha , STD ,and the RMWSD statistics. These dif statistics were computed for each of the nine sample size conditions and each test. The results were then compared to the corresponding asymptotic dif indices obtained in the real data by computing rank order correlation coefficients between each dif statistic's value and the asymptotic index for the same items. Agreement between the asymptotic index and the dif statistic on the items flagged as exhibiting dif were also computed. The process of simulating the tests, generating the responses, and comparing asymptotic index and dif statistics was replicated 100 times.

The correlation studies indicated that the MH-alpha asymptotic index correlated highly with the MH-alpha dif statistics. Similar results were noted for STD statistic and STD asymptotic index. However, this was not true with the RMWSD index. For each of the three dif statistics, the correlation with the corresponding asymptotic index diminished with the decrease in sample size.

The MH-alpha statistic underestimated its asymptotic index for sample sizes of 2000 but overestimated the asymptotic index as the sample size decreased. The STD statistic overestimated its asymptotic index, regardless of sample size. Similar trends were observed with RMWSD index. The MH-alpha statistic and STD statistic correctly identified biased items at similar rates for samples equal to or greater than 250. For smaller sample sizes the STD statistic had a higher false positive rate than MH-alpha

statistic. The RMWSD statistic had very high false positive rates across small and large sample sizes. As sample sizes decreased, the three indices showed increased variability. The MH-alpha statistic produced unbiased results for moderate sample sizes and moderate test length. It was concluded that the MH-alpha was a more viable index for dif detection and analysis.

In the analysis and classification of dif items in test data using dif statistics, tests of significance level or cutoff scores are used. Tests of significance are appropriately used only when the dif index has a known associated statistical test and distribution. However, in instances where no known associated statistical test exists and the distributions of the indices are not known, cutoff scores in terms of the standard deviation of the indices or percentiles are used.

Sykes and Fitzpatrick (1990) conducted a study to examine a multiple-method in which MH-alpha cutoff scores in identifying and classifying dif items were combined with an IRT based (using Rasch item difficulty) cutoff score so as to establish a maximised decision consistency in identifying and classifying dif items.

Subjects were 68,458 examinees who sat for a professional licensure examination (CTB/MacGraw Hill, 1988). The test consisted of 299 items. For analysis, examinees were classified into eight categories of ethnic groups. Ethnic group one had 47,573 members, group two had 6486, group three 5466, group four 2004, group five 1014, group six 486, group seven 307, group eight 746. Some 4396 examinees were not classified into any category and were therefore not used in the analyses of MH-alpha and IRT methods in the study.

Two analyses of the MH-alpha were carried out in which total test scores were used as matching variables. In the analysis total test scores were divided into 13 levels such that each category had at least 50 examinees for reference and focal groups. Using ethnic group one as the reference group, comparisons were made between group one and two, group one and three, group one and four and finally group one and five. In the second analysis of MH-alpha, total scores were classified into nine categories with at least 22 examinees per category. Ethnic group one was then compared to groups five, six, seven, and eight. Since groups six, seven, and eight were small, fewer categories were used in this analysis so as to ensure that there were enough examinees in each score category for adequate matching.

For the analysis of the IRT methods, 500 examinees were randomly sampled from ethnic group one. A random sample of 1000 examinees was drawn from each of group two, three, and four. All examinees in groups five, six, seven, and eight were used in the analysis as these groups were very small. The IRT method used in the study was based on the IRT procedure suggested in Lord's study (Sykes & Fitzpatrick 1990) of estimating item difficulty. First, the Rasch item difficulty, 'b' parameter, for each item was estimated using LOGIST and a sample of 500 examinees was randomly drawn form the reference group. Rasch ability estimates for these examinees were generated. Item difficulties were then re-estimated each time for each comparison of reference and focal groups. The resulting item difficulties were rescaled to place them on the same scale of reference and focal groups on item performance. A t-statistic for each item in each comparison was calculated. This was to determine the difference between item difficulty

for the reference and focal groups.

For each reference-focal group comparison, the MH-alpha estimates obtained for each item were plotted on a bivariate plot against the t-statistic obtained for the same item. Selected cutoff scores of the MH-alpha were represented on the y-axis; perpendicular lines to the y-axis passing through the cutoff scores were drawn. Similarly, for the t-statistic cutoff scores were represented on the x-axis and perpendicular lines to the x-axis were drawn for selected cutoff scores. The intersection of these lines created four quadrants.

Quadrant 1, contained items with MH-alpha and t-statistics that were greater than or equal to the selected MH-alpha cutoff score and the t-statistic cutoff score. These items were classified as potentially biased. Quadrant 3 contained items with both MH-alpha and t-statistics less than the selected MH-alpha and t-statistic cutoff score. Quadrant 1 and 3 therefore contained items consistently classified on the basis of the two cutoff scores. Quadrant 2 and 4 contained items that were inconsistently classified on the basis of the two indices cutoff scores.

The total number of items found in quadrant 1 and 3 were expressed as a proportion of the total number of items analyzed. This was done for each combination of cutoff scores selected for MH-alpha and t-statistics. The proportion was regarded as a measure of consistency between MH-alpha and t-statistic and termed as concordance proportion. By selecting different MH-alpha cutoff scores and t-statistic cutoff scores and counting the items in quadrant 1 and 3, each plot was systematically searched using a computer algorithm to find the combination of MH-alpha and t-statistics cutscore values

that produce the largest (maximum) count of items. These pairs of cutoff scores were considered to have produced the maximum concordance proportion for each comparison. The Kappa statistic (k-statistic) was also computed as a measure of consistency.

Results indicated that the concordance proportions were large, ranging from 0.94 to 0.99. This implied a high level of consistency for the two indices. Correlations between MH-alpha and the t-statistics for each comparison were also high, suggesting a strong positive relationship between values produced by MH-alpha and the IRT method. The cutoff scores multiple-method was compared to results obtained by the traditional significance level, using chi square tests for the MH statistic across all the reference-focal groups comparisons. Results indicated that under the chi-square test of significance, the number of flagged dif items in each comparison depended on sample size while in the case of cut scores multiple-method, the number of flagged dif items was not related to sample size. It was also observed that lower cutoff scores were derived from those comparisons with lower and less variable levels of dif while higher cutoff scores were derived for those comparisons that appeared to have items with large amounts of dif.

The use of multiple methods enabled the detection of potential dif items by two independent methods. This gave these methods an advantage over the level of significance test approach, in terms of accuracy and generalizability.

Raju, Bode, and Larsen (1989) carried out a study to empirically evaluate the effects of the number of score groups and the inclusion or exclusion of studied dif items in forming score groups on the MH-alpha index.

A 40 item test was administered to 3795 examinees (2400 Whites, 1161 Blacks

and 234 Hispanics). Two comparisons were made between Whites and Blacks as well as Whites and Hispanics. In each case Whites formed the reference group. For each of the comparisons, 10 dif analyses were conducted.

In the first analysis, using total scores on the test as the matching variable, white and black examinees were each divided into two score groups. The first score group consisted of examinees with scores from zero to 20. The second score group consisted of examinees with scores from 21 to 40. In the second analysis of two score groups, similar classification was done but the score for the studied item was excluded. The next analysis was done using four score groups with cutoffs at 11, 21 and 40 and separately analyzing the data for MH-alpha estimates by including the studied item as well as excluding the studied item. Similar analyses were conducted using six score groups, eight score groups and 10 score groups. This resulted in different analyses of MH-alpha estimated, with two analyses for each of the five score groupings. Intercorrelation of MH-alpha estimates across different score groups were made for each comparison.

Results indicated that four score groups and more, yielded better MH-alpha estimates than the two score groups. The intercorrelation of MH-alpha across different score groups indicated that four score groups correlated highly with the higher score groups of six,eight, and ten. Higher consistency was noted in correct dif item identification when four score groups or more were used than when two score groups were used. Including the studied dif item score yielded better MH-alpha estimates than excluding the studied item. As the number of scores increased the difference in the two cases diminished. The results of this study differed with those from Wright's study

(Raju, Bode & Larsen, 1989). The explanation provided was that optimal score groups may be a function of the range of total scores used (Raju, Bode, & Larsen 1989).

Currently, relevant literature on the logistic regression procedure (LR) in detecting dif items and analysis of the distribution is rather limited. Even more so are the comparative studies between the MH -procedure and LR procedure. Swaminathan and Rogers (1990) conducted a study to compare the accuracy and power of the MH procedure with that of the LR procedure in detecting uniform and nonuniform dif items. It was shown that the MH procedure can be represented as a logistic regression model with a discrete ability variable and no interaction term between ability and group membership variables. For the LR procedure, the logistic regression model is stated with a continuous ability variable and an interaction term between ability and group membership.

So as to carry out the comparison between the two procedures, simulated data were used so that tests with known dif conditions could be studied. The following factors were manipulated in the study; sample size with two levels (250 per group and 500 per group), test length with three levels (40 items/60 items/80 items) and the nature of dif with two levels (uniform and nonuniform dif). These factors were hypothesized to have an effect on the power of the two procedures.

Within each test, 20 percent of the items were induced with dif such that half of the dif items were uniformly biased, while the other half were nonuniformly biased. In simulating uniform dif the item difficulty parameter 'b' was varied. For nonuniform bias, the discrimination parameter was varied.To generate a specified amount of dif, the

item parameters were selected such that the pre-specified area between item characteristic curves was computed using the formula given by Raju (Swaminathan & Rogers, 1990). Item responses were then generated with a three parameter item response logistic model, using the DATAGEN program (Hambleton & Rovinelli, 1973).

The two dif detection procedures were compared with respect to the percentage of items with dif that were correctly identified. Those items without dif were analyzed for false positive rates. In addition to this the power of each procedure (MH and LR) was analyzed by carrying out twenty replications of the 80 item test and 500 examinees. This represented the longest test and the largest sample size.

Results showed that the MH procedure and the LR procedure were equally effective in detecting uniform dif with 75 percent accuracy at the sample size level of 250. At the sample size of 500, both procedures detected uniform dif with 100 percent accuracy. For nonuniform dif the MH procedure was unable to detect dif under any of the conditions, but LR detected nonuniform dif with an accuracy of 50 percent in the small sample size of 250 and in short tests. With large sample sizes of 500 and a long test,the LR procedure detected nonuniform dif with about 75 percent accuracy.

The false positive rates for the MH procedure were lower than those of the LR procedure in both dif conditions. It was concluded that the MH-procedure is as powerful as the LR procedure in detecting uniform dif but LR procedure is more powerful in detecting nonuniform dif .

In their second study to compare LR statistics with the MH procedure for detecting dif, Rogers and Swaminathan(1990) simulated 16 conditions for each

group.They manipulated the following conditions; two levels of test length (40 items/80 items), two levels of data fit (two parameter logistic model and three parameter logistic model), two levels of score distributions (normal and skewed), and two levels of proportion of biased items.

Two simulations were carried out for the sample sizes of 250 and 500. The first simulation was used to examine the distribution of the test statistic of the LR and the MH procedures. For the procedures to be effective in dif detection, they were expected to satisfy the distributional assumptions on which they were based. The second simulation was used to investigate the power of the two procedures.

It was hypothesised that sample size, test length, model data fit, score distribution and proportions of biased items would affect the power and accuracy of the LR and MH statistics. The simulated data were generated using the two parameter logistic model (good fit) and three parameter logistic model (poor fit). In the three parameter model the 'c' parameter was fixed at 0.2. Dif items were generated and the amount of bias specified using the area method based on Raju's formulae.Uniform dif was simulated by varying the 'b' parameters. Nonuniform dif was simulated by varying the 'a' parameter.

For each of the five selected items with various levels of difficulty and discrimination, data were generated for two groups using the same parameters for each group. One hundred replications of the data were conducted for LR and MH statistics. An empirical sampling distribution was then constructed. The Kolmogorov-Smirnov test was performed to determine if the test statistics (LR and MH) had the expected distributions. The results showed that at the 500 sample size, the observed distribution

was closer to the expected distribution, compared to the case of the 250 sample size in which the observed distribution was not as close to the expected distribution as in the first sample.

In the power study to determine the effectiveness of LR and MH statistics in detecting uniform and nonuniform dif, 35 dif items were constructed (16 uniform dif and 19 nonuniform dif items). Items without dif were generated using item parameters values taken from real data and chosen to produce normally distributed ability values of both groups. Items with dif were incorporated one at a time into each data set, then removed after the dif statistic was calculated. Twenty replications of each condition were conducted. Percentage of uniform dif and nonuniform dif detected by each procedure was computed for the proportion biased (15 percent bias). ANOVA was conducted for the effects of all factors stated on the performance of the LR and MH procedure in detecting uniform dif and nonuniform dif. The dependent variable was how frequently dif items were flagged out of twenty replications. Separate analyses were done for the detection of uniform and nonuniform dif.

The detection rates for both indices increased with an increase in sample size. Both LR and MH indices were not affected by the shape of the score distributions and the proportion of biased items. While LR was affected by test length the, MH statistic was not affected. However, both of the indices were affected by the model underlying the data (data fit). The detection rate of LR was lower for the three parameter logistic model than for the two parameter logistic model. For the MH statistic the detection rate also dropped significantly when the three parameter logistic model was used instead of

two parameter model.

For both indices items of moderate dif and high discrimination were more easily detected than items of high difficulty and those with low discrimination.Higher amounts of dif were more easily detected by the two indices. MH and LR indices detected uniform dif at about the same rate. However, MH could not detect nonuniform dif at the same rate as LR statistic (57 percent detection rate for LR and 34 percent for MH for 3 parameter model data). It was also noted that for low difficulty items, the MH statistic had a lower detection rate than LR. With high difficulty items the detection rates were the same for uniform dif. The most significant findings in the study, were the effects of sample size and item type (uniform and nonuniform dif) on the detection rates of MH and LR procedures.

Rogers and Swaminathan(1990) provided an explanation of the fact that MH was not effective in detecting nonuniform dif, in terms of interaction between ability and group membership.The MH statistic is sensitive to the main effect of group membership. Since this can be detected in the presence of ordinal interaction, MH statistic can detect only this type of nonuniform dif.

Summary of findings

Sample size was found to affect the performance of MH statistics in that, with the increase in sample size, the detection rate also increased (Mazor, Clauser & Hambleton, 1991; Gutierrez, 1989; Spray, 1989; Swaminathan & Rogers, 1990;Rogers & Swaminathan, 1990). Similar results were noted for the LR statistic (Swaminathan &

Rogers ,1990; Rogers & Swaminathan,1990). However, with a decrease in sample size, the detection rates of these indices dropped.

Poorly discriminating items and items with high difficulty were less likely to be detected by both MH and LR statistics except at large sample sizes. However items of moderate difficulty and high discrimination were most likely to be correctly identified by the MH indices (Mazor, et.al., 1991; Gutierrez 1989; Swaminathan & Rogers, 1990; Rogers & Swaminathan,1990) and the LR statistic (Swaminathan & Rogers,1990;Rogers & Swaminathan,1990).

Test length affected the detection rate of the MH procedure in that test scores were used as the matching variable and longer tests provided a more reliable matching criterion (Donoghue & Allen, 1991). The number of score groups (Raju, Bode, & Larsen, 1989), and the matching or pooling method of scores also affected the performance of the MH procedure (Donoghue & Allen, 1991). In the case of LR statistics since total test score is used as a predictor in the statistic, a reliable test score which is obtained from long tests was shown to yield better results in terms of detection rates for both uniform dif and nonuniform dif (Swaminathan & Rogers, 1990; Rogers, & Swaminathan, 1990). Simulation studies have also shown that LR is more powerful than the MH procedure in the detection of nonuniform dif.

MH indices were also examined under equal ability distributions for the reference and focal groups and under unequal ability distributions for the groups (Mazor et. al. 1991). When ability distributions of the reference and focal groups were equal, correct identification rates were higher than when the ability distributions were unequal for the

two groups.

While the distribution of MH-CHISQ and LR statistics are known to be chi-square distributions with means of one, standard deviation of 1.41 and skewness of 2.82,the mean of MH-Z is zero and the standard deviation is not known. Gutierrez (1989) studied the influence of sample size, 'a' parameter, and 'b' parameter on the distribution of the MH-Z index. From the significant effects of these variables on the standard deviation of MH-Z, it was deduced that the MH-Z index is unstable across levels of these variables and this in turn affected the power and distribution of this index.

Studies of distributions and characteristics of dif indices where true dif conditions are known have been conducted using simulated data. Hambleton and Rogers (1989) showed that the simulated data very closely approximate results from real data. This provides a sound rationale in using simulated data for dif analysis and studies.

MH indices have been compared to other dif indices (Camilli & Smith, 1990; Spray, 1989; Swaminathan & Rogers, 1990; Sykes & Fitzpatrick, 1990) and were found to be in agreement with the indices except in the identification of nonuniform dif. In identifying uniform dif the MH procedure was found to be very powerful especially with large sample sizes, high 'a' values, and moderate 'b' values.The LR statistics (Swaminathan & Rogers,1990; Rogers & Swaminathan, 1990) were found to be effective in detecting nonuniform dif with a relatively high accuracy, at large sample sizes.

Differences in ability distributions between the reference and focal groups were noted to influence the power of the MH procedure although little is known on the

possible effects on LR statistics. This is an area for further research. Cutoff scores established to identify dif at P_{95} and $P_{97.5}$ (Gutierrez 1989) influenced the detection rates of MH-Z, and LR statistics (Rogers & Swaminathan,1990) at the P_{99} and P_{95}. However little is known about the effects of sample size, 'b', and 'a' parameters on the cutoff scores of the LR statistics. The effects of 'a' and 'b' parameters at various sample sizes on the performance and distribution of MH and LR statistics are areas for further research. These were investigated in this study. In this study simulated data are used.

A justification of the use of simulations is that most MH research has been done using real data where true state of dif is not known. No logical conclusions can be made in such studies. So far, research done where simulation has been used and replications done, have been limited in number.

Other than studies by Guiterrez (1989) on the distribution of the MH-Z index, and the study by Swaminathan and Rogers (1990) on the distribution of the MH-CHISQ and LR using limited replications, there have been few studies on the distribution of the indices.Therefore, there is a need to examine the distribution of the indices of MH and LR as well as variables that might affect these distributions.

Purpose of the Study

The purpose of this study was to examine the distribution and characteristics of the four indices, MH-delta, MH-CHISQ, and LR test statistics (for uniform, LU and nonuniform,LN) under controlled conditions of the null hypothesis, using simulated data. Under the null condition the indices are not expected to flag any item as exhibiting dif,

except by chance or due to random error, irrespective of the data. Different controlled conditions were simulated under the null hypothesis of no dif, in order to examine their effects on the distributions of the MH-delta, MH-CHISQ, LU and LN test statistics.

While the MH-delta has an advantage over the other MH-statistics in that it provides a measure of magnitude and direction of the dif in a studied item, the MH-CHISQ has an advantage over the other MH-statistics in that it has a known test of significance, a chi-square test with one degree of freedom.

The LR test statistics can be tested for significance using a chi-square test with two degrees of freedom, simultaneously for uniform and nonuniform dif. However, it is also possible to test for significance for uniform dif and nonuniform dif independently using two chi-square tests with one degree of freedom each.The one degree of freedom test is expected to be more powerful than the case where uniform and nonuniform dif are tested for significance simultaneously using the two degrees of freedom chi-square test.

In this study the effects of sample size, 'a' parameter, 'b' parameter, and differences in ability distribution between reference and focal groups, were examined.

Research Questions

This study is an experimental research . However, it remains exploratory because it is not possible to predict with any certainity, what will happen, with the exception of sample size. For sample size, it is expected that the chi-square values will increase with an increase in sample size.

The study attempted to answer the following research questions:

1) Are the distributions of MH-CHISQ, MH-delta, LU, and LN affected by sample size, value of 'a', value of 'b', and equality of distribution of focal and reference groups?
2) Are the cutoffs at P_{90} and P_{95} of the indices affected by changes in the stated variables independently or by a combination of these variables? That is, are the main effects or interaction effects of the variables significant on the distribution of the indices?
3) Are there differences in the observed distributions of the indices obtained, compared to the expected distributions?

These research questions were answered through the study design given in the next chapter.

CHAPTER III

METHODOLOGY

The main concern of the study was to evaluate the effects of sample size, 'a' parameter, 'b' parameter, and differences in ability distribution (between the reference and focal groups) on the distributions of MH-statistics (MH-delta and MH-CHISQ) and LR(LU and LN) statistics for uniform and nonuniform dif.

The study was conducted under the null condition of no dif. The methodology of the study is in five sections. The first section consists of the data collection approach while the second section consists of the assumptions made. The other sections consist of characteristics of the items, characteristics of examinees, simulation model and data analysis procedures.

Data Collection Approach

A Monte Carlo study was conducted instead of utilizing real data. This method was chosen because in simulation studies the true state of dif is known. Simulation studies also allow the researcher to manipulate the variables of interest. Replications can then be made to evaluate the stability of the results.

The DATAGEN computer program (Hambleton & Rovinelli,1973;Carlson 1983) was used to generate item response strings with item parameters set to specified values for both focal and reference groups. The abilities were drawn from normal distributions for both groups. The values of MH-statistics were computed using the MANTEL program (Ackerman, 1987). The LR test statistics were computed using an estimation program developed by Spray (1989b).

One hundred replications for each combination of condition described in the research design were made for each dif index(See Table 2).SPSSX programs were used to analyse the data.

Table 2

Conditions under which the four Indices were Studied Using a 63 Item Test for each of the two levels of Ability Distribution over 100 Replications.

	Sample size 300/100 'a' values of;			Sample size 600/200 'a' values of;			Sample size 1200/400 'a' values of;		
'b' values	0.6	1.0	1.4	0.6	1.0	1.4	0.6	1.0	1.4
-2.0									
-1.8									
-1.6									
-1.4									
-1.2									
-1.0									
-0.8									
-0.6									
-0.4									
-0.2									
0.0									
0.2									
0.4									
0.6									
0.8									
1.0									
1.2									
1.4									
1.6									
1.8									
2.0									

Low 'b'=-2.0 to -0.8

Medium 'b'=-0.6 to +0.6

High 'b' =+0.8 to +2.0

Assumptions Made in the Study

It was assumed that the tests were unidimensional. It was also assumed that the number of items used and the resulting test scores are sufficient and unbiased estimators of the ability of interest. Although ability is a continuous variable, it was assumed that the dichotomously scored items would approximate the trait without significant loss of information. In the study, examinees were assumed to have reached and responded to all test items.

Characteristics of the Items

The simulated test consisted of 63 items. This test length was chosen for high reliability since long tests are more reliable than short tests. In MH and LR test statistics, total test scores are used as ability measures while in MH-statistics reliable test scores provide improved matching criterion.

Item response strings were simulated using a three parameter logistic model. The 'c' parameter was set at 0.15. Parameters 'a' and 'b' were set at predetermined values. Three levels of the item discrimination parameter 'a' were used; (a = 0.6, a = 1.0, and a = 1.4). For a given test all items had the same 'a' value.The three 'a' values were chosen to evaluate the effect of a small 'a' less than one, at 'a' equal to one and at 'a' greater than one. For each value of 'a', a 63 item test was simulated.

The item difficulty parameter 'b' was set with 21 levels ranging from -2.0 to +2.0 , at intervals of 0.2. Since there were 63 items, for each 'b' value interval of 0.2 (between -2

to +2) there were three items.

Characteristics of Examinees

In order to study the effects of sample size on the distribution of MH-statistics (MH-delta and MH-CHISQ) and LR(LU and LN) test statistics three different sample sizes were simulated. Sample sizes of 400, 800 and 1600 were studied. In each sample, the ratio of focal group to reference group was set at 1:3. This is the ratio that has been used in several dif studies, although even smaller ratios have been used before (Spray 1989; Phillips & Mehrens, 1988). In the first sample of 400, the focal group consisted of 100 examinees while the reference group were 300. In the second sample of 800, the focal group consisted of 200 examinees while the reference group were 600. The third sample consisted of 1600 examinees of whom 400 were focal group members and 1200 were reference group members. These three sample sizes were selected such that the increase in sample sizes in multiples could assist in evaluating any linear trends that might exist in the effects of the variables on the distributions of the MH and LR test statistics.

The ability distributions of the examinees were studied in two levels. The first level was when the ability distributions of the focal and reference groups were equal. The mean of each group was set at zero and a standard deviation of 1.0. For the second level the two groups' ability distributions were different. The reference group's mean was set at zero and the standard deviation 1.0, while the focal group's mean was set at -0.5 and the standard deviation 0.83. The distributions of the four indices were evaluated under the two cases of equal and unequal ability distribution.

Simulation Model and Data Analysis Procedure

Using a 63 item test, as described above, for each of the conditions of three sample size by three levels of 'a' by two levels of ability distributions, data were simulated. The 21 values of 'b' generated , were divided into three categories of low 'b' ,medium 'b' and high 'b'.The low 'b' ranged from -2.0 to -0.8, medium 'b' ranged from -0.6 to +0.6, and high 'b' ranged from +0.8 to +2.0.

The four indices were computed and relevant descriptive statistics(mean, standard deviation, kurtosis and skewness) obtained for each index in each category of 'b'. Cutoffs at P_{90} and P_{95} were also obtained for each index, in each category of 'b'.The process was repeated for one hundred replications . For the MH-delta index, absolute values were used in determining the two percentiles.

Analysis of the indices was done in two stages. The first one was on the distribution of the indices while the second one was on the percentile cutoffs.

Analysis of the Effects of the Variables on the Distributions of the Indices.

To assess the effects of sample size , ability distribution, discrimination, and item difficulty as the independent variables a MANOVA was conducted on each of the four indices (MH-delta, MH-CHISQ, LU,and LN) using the descriptive statistics(mean, standard deviation, skewness, and kurtosis) as dependent variables. Levels of significance for post hoc procedures were set at 0.05 for MANOVA and 0.01 for ANOVA. When significant effects occured, univariate ANOVAs were conducted.

Analysis of False Positive Rates.

Using the four independent variables a MANOVA was conducted on the two percentiles (P_{90},and P_{95}) for each index. When significant effects occured ,then univariate ANOVAs were conducted to determine whether the 'a' parameter, 'b' parameter, sample size, and ability distribution had effects on the cutoffs at the stated percentiles.

CHAPTER IV

RESULTS AND DISCUSSION

The results of this study are presented and discussed in this chapter. So as to analyse the effects of the independent variables on the distribution of the four indices, that is MH-CHISQ, MH-Z, LU and LN, a separate MANOVA was conducted on the descriptive statistics for each of the indices (mean, standard deviation, skewness, and kurtosis).

So as to analyse the effect of the independent variables on the cutoff values for the false positive rates of 0.10 and 0.05, a MANOVA was conducted on P_{90} and P_{95} for each of the four indices.

For each multivariate analysis, significance was taken at $p < 0.05$, using the Pillai's statistic as it is robust to violation of the assumption of homogeneity of variance-covariance matrices. Each significant multivariate result was followed by a further univariate ANOVA at $p < 0.01$ for the mean, standard deviation, skewness , kurtosis, and the two percentiles. This level of significance was chosen so as to control type I error rate.

In the univariate results, significant four-way interaction and significant three way interactions occurred. Due to difficulties in the interpretation of four way and three way interactions, separate post-hoc one-way ANOVAs were conducted for each independent variable that contributed to the significant four-way or three-way interaction. If two-way interactions were found, simple effects were carried out. For each

of the of the one-way ANOVAs ,if significance was found at 0.01 level, Scheffe tests were conducted to identify differences between pairs. Where significant main effects were present but not in two-way interactions, Scheffe tests were carried out at 0.01 level of significance. For all main effects Scheffe tests were also conducted at the 0.01 level. The results are analysed for each index at a time.

Effects of the Independent Variables on the Distribution of MH-CHISQ.

Following a significant multivariate effect, univariate ANOVAs were conducted. The MANOVA and univariate results on the mean of MH-CHISQ showed a multivariate and univariate significant two-way interaction effect of discrimination by item difficulty. Sample size was found to be a significant main effect in the analysis of the mean of MH-CHISQ, over replications. Scheffe's post-hoc pairwise comparisons on the levels of sample size was conducted at 0.01 level of significance. All pairs of the groups were found to be significantly different.

At sample size of 400, the group mean was 0.758, at sample size of 800 the group mean was 0.835, while at the sample size of 1600, the group mean was 0.877. Increase in sample size resulted in a mean increase in MH-CHISQ.

So as to analyse the significant two-way interaction effect of discrimination by item difficulty on the means of MH-CHISQ, simple effects analyses were conducted. Table 3 shows the group means for the MH-CHISQ at each discrimination and item difficulty level, as well as the post-hoc Scheffe results.

Table 3

Group Means of MH-CHISQ Means at Discrimination by Item Difficulty two-way Interaction.

Discrimination level	Item difficulty level			Difference between groups
	Low	Medium	High	
a_1 = 0.6	0.822	0. 836	0.854	N.S
a_2 = 1.0	0. 755	0.857	0.852	(1,3)(1,2)S
a_3 = 1.4	0.7445	0.842	0.829	(1,3)(1,2)S
Difference between groups	(1,3)S	N.S	N.S	

S = Significant at 0.01 level

N.S =Not significant for one-way ANOVA

Scheffe's test conducted at 0.01 level of significance showed that at low item difficulty the group mean of discrimination at 0.6 was significantly higher than the mean of discrimination at 1.4. There were no other significant effects.

Across discrimination level of 0.6, there was no significant difference found for item difficulty. At discrimination levels 1.0 and 1.4, the mean at low item difficulty was significantly smaller than for medium and higher item difficulty,. From Table 3, it is evident that there is a disordinal interaction between levels of discrimination and item difficulty.

It can be deduced that, with an increase in sample size, the means of MH-CHISQ index also increased. The effects of item difficulty and discrimination were significant over selected levels of the other variable with item difficulty showing more differences.

For the standard deviation of the MH-CHISQ index there was a multivariate and univariate significant two-way interaction of discrimination by ability distribution. Sample size and item difficulty showed significant main effects.

Scheffe's test was conducted for sample size. The means of the standard deviations of MH-CHISQ at sample size of 400 (m=1.1310) was significantly smaller than that of 800 (m=1.1965) and that of 1600 (m=1.2194). However the means of standard deviations at sample size of 800 and 1600 did not differ significantly.

The Scheffe test was also conducted for item difficulty as a significant main effect on the standard deviation of MH-CHISQ index. The mean of the standard deviation at low level difficulty (m=1.148) was significantly lower than medium (m=1.20) and high (m=1.19).

The two-way interaction of discrimination by ability distribution for the standard deviation of MH-CHISQ index was followed up by a simple effects analysis so as to determine which levels of discrimination and ability distribution differed significantly at 0.01 level of significance (see Table 4).

There were no significant differences for ability at any level of discrimination. The only significant difference for discrimination was between discrimination level 0.6 (m=1.218) and level 1.4 (m=1.130) for unequal ability (see Table 4).

Table 4

Group Means of the Standard Deviation of MH-CHISQ for the Discrimination by Ability Distribution Interaction

Ability Distribution	Discrimination level			Difference between group
	0.6	1.0	1.4	
Equal ability	1.165	1.203	1.192	N.S
Unequal ability	1.218	1.182	1.130	(1,3)S
Differences between groups	N.S	N.S	N.S	

N.S = not significant, one-way ANOVA

S = significant at 0.01 level

Scheffe's test conducted at 0.01 level of significance.

It can be deduced that sample size and item difficulty had significant effects on the standard deviation of MH-CHISQ. Discrimination and ability distribution only had effects on the standard deviation of MH-CHISQ for discrimination at unequal level of ability.

The MANOVA and ANOVA results on the skewness showed that sample size was found to be the only significant effect. So as to analyse the effects of sample size on the skewness of the distribution of MH-CHISQ, a Scheffe test was conducted. Scheffe's test of pairwise comparison showed that the group mean of skewness at low sample size of

400 was significantly higher than those of high sample size of 1600. Similarly the group mean of skewness at low sample size of 400 significantly differed from that of medium sample size of 800. There was no significant difference between medium and high sample size. The group mean at sample size levels were; low sample size (m=0.477) at medium sample size (m=0.429) and at large sample size (m=0.415). As expected the group means of the skewness were positively skewed. This is a characteristic of a chi-square distribution.

In the case of the kurtosis of the MH-CHISQ index, sample size was also found to have a significant main effect for MANOVA and ANOVA. This was followed by a Scheffe test, so as to analyse the effect of sample size on the kurtosis values of MH-CHISQ. Scheffe test showed that a low sample size of 400 and high sample size of 1600 were significantly different. The means of the kurtosis at sample size levels were as follows; at low sample size of 400, the group mean was 4.842 while at medium sample size of 800, the group means of kurtosis was 4.419. At high sample size of 1600 the group mean of the kurtosis was 4.279. At smaller sample size the distribution was more 'peaked' than at high sample size. That is, the distribution of MH-CHISQ was leptokurtic while at high sample size level the distribution was less leptokurtic.

It can be concluded that the distribution of MH-CHISQ index is influenced by sample size for all the stated descriptive statistics, and item difficulty for the standard deviation . Ability distribution had no significant main effects on any of the analysed descriptive statistics other than through an interaction with discrimination for the standard deviation of MH-CHISQ. Even then, at no level of discrimination were group means of the

standard deviation significantly different between ability distribution levels of equal and unequal distribution.

Effects of Independent Variables on the P_{90} and P_{95} of MH-CHISQ.

A MANOVA was conducted on the two percentiles (P_{90} and P_{95}) of MH-CHISQ so as to analyse the effects of the independent variables. The P_{90} corresponds to the cutoff for the 0.10 false positive rate, while P_{95} corresponds to the cutoff value for the 0.05 false positive rate. At P_{90} of MH-CHISQ two significant multivariate and univariate main effects occurred, namely sample size and item difficulty. A two-way interaction between item difficulty and ability distribution also occurred for P_{90} of MH-CHISQ.

To determine the effects of sample size as a significant main effect the Scheffe test of pairwise comparison was conducted. Results showed a significant difference which was between the low sample size of 400 (m=2.572) and medium sample size of 800 (m=2.758). Similarly there was a significant difference between sample size of 400 and that of 1600 (m=2.840). However there was no significant difference between the means of P_{90} at sample size 800 and those of 1600. At sample sizes of 800 and 1600, the means of P_{90} were greater than the expected value of 2.70, which is the tabled value at 0.10 level of significance for a chi-square distribution with one degree of freedom. This implied that at sample sizes of 800 and 1600 one would obtain higher false positive identification of biased items when the tabled value was used instead of P_{90}. The contrary is true when computing false positive rates with P_{90} of MH-CHISQ at the low sample size .

Item difficulty also contributed in a multivariate and univariate significant two-way interaction of ability distribution by item difficulty. For this reason the significant main effects of item difficulty will not be analysed.

The two-way interaction was analysed by conducting a simple effects analysis. Table 5 shows the group means of P_{90} of MH-CHISQ for the interaction of ability distribution by item difficulty.

Table 5

Group Means of the P_{90} of MH-CHISQ for the two-way Interaction of Ability Distribution by Item Difficulty

Ability distribution	Item difficulty			Group difference
	Low	Medium	High	
Equal ability	2.580	2.853	2.748	(1,3)(1,2)S
Unequal ability	2.640	2.711	2.795	(1,3)S
Group difference	N.S	N.S	N.S	

N.S = Not significant one-way ANOVA

S =significant at 0.01 level

Scheffe test conducted at 0.01 level of significance.

From the group means in Table 5 the two way significant interaction of ability by item difficulty is a disordinal interaction. From Scheffe's test of pairwise comparisons at equal ability distribution, the group means at low item difficulty are less than those at medium and high level of item difficulty. At unequal ability distribution, the low item

difficulty mean was significantly less than for high item difficulty. For each item difficulty level, the pairwise comparison for ability distribution was not significantly different. It can be deduced that ability moderates the effect of item difficulty. It is observed that averaged on ability distributions, the P_{90} of MH-CHISQ differed significantly across item difficulty. Item difficulty has therefore, a significant effect on the stability of the P_{90} of MH-CHISQ, but slightly different effects at two ability levels.

For P_{95} of MH-CHISQ, sample size occurred as a significant main effect. A multivariate and univariate significant two-way interaction effect of discrimination by ability distribution also occurred for P_{95}.

In order to analyse the effect of sample size on P_{95} of MH-CHISQ, the Scheffe test was conducted. Scheffe test of pairwise comparisons of the levels of sample size showed that there was a significant difference between low sample size of 400 (m=4.125) and the medium sample size of 800 (m=4.315), as well as low sample size of 400 and high sample size of 1600 (m=4.432). All the three means of P_{95} at the sample size levels were above the expected value of 3.84 (tabled), for a chi-square distribution with one degree of freedom at 0.05 level. The use of the tabled values would result in higher false positive rates than those found using the 0.05 level from the data.

The significant two-way interaction effect for P_{95} of the MH-CHISQ index of discrimination by ability distribution was analysed by conducting a simple effects analysis (see Table 6).

Table 6

Group Means of the P_{95} of MH-CHISQ for the two-way Interaction of Discrimination by Ability Distribution

Ability distribution	Discrimination level 'a'			Group difference
	0.6	1.0	1.4	
Equal ability	4.216	4.386	4.334	N.S
Unequal ability	4.483	4.295	4.101	(1,3)S
Group Difference	N.S	N.S	N.S	

S = significant at 0.01 level

N.S =Not significant one-way ANOVA

Scheffe's test conducted at 0.01 level of significance.

From Table 6, it is observed that there are no significant effects on each discrimination level across ability distribution. At unequal ability distribution, discrimination level of 0.6 (m=4.483) had a mean greater than that of discrimination level 1.4 (m=4.101). The difference for unequal ability distribution was possibly a chance result.

Significant effects of the independent variables on the distribution of MH-CHISQ were almost the same as for the percentiles. For all six dependent variables (mean, standard deviation, skewness, kurtosis, P_{90}, and P_{95} of MH-CHISQ) sample size

was observed to be a significant main effect. Item difficulty affected the distribution of MH-CHISQ as a significant main effect for the standard deviation of MH-CHISQ. The mean of MH-CHISQ was affected by item difficulty through the interaction effect with discrimination. In the case of the percentiles, item difficulty had a significant main effect on the P_{90} through interaction with ability distribution. However item difficulty had no significant effect on P_{95} of MH-CHISQ.

From the results of the effects of the independent variables on the standard deviation of the index,it can be deduced that MH-CHISQ is unstable across sample size and item difficulty. Discrimination variable affected the P_{95} through the interaction effect with ability distribution.

The results of these findings agree with Camilli and Smith(1990) in which it was found that MH-CHISQ was significantly affected by sample size. Table 7 is a summary of the descriptive statistics of the independent variables for the MH-CHISQ index. The results can be compared with the expected (Tabled) values.

From Table 7 the means of the MH-CHISQ were consistently under estimated in relation to the tabled value of 1.0. Similar results were found with standard deviations and skewness. The standard deviation of MH-CHISQ expected was 1.4 while the expected skewness was 2.82. However, the skewness values were positive, which is expected of a chi-square distribution. The kurtosis values reflected a leptokurtic distribution, that is a highly 'peaked' distribution.

The overall means of the P_{90} at low item difficulty, high discrimination, and at low sample size were all below the tabled or expected value of 2.70, and therefore

underestimated.

Table 7

Descriptive Statistics of MH-CHISQ, over 100 Replications across the Independent Variables.

	Mean	Std	Skewn	Kurt	P_{90}	P_{95}
Sample Size						
300/100	0.758	1.131	2.0851.	4.842	2.572	4.125
600/200	0.853	1.197	1.996	4.420	2.578	4.351
1200/400	0.877	1.219	1.961	4.279	2.837	4.432
Discrimination						
a_1 = 0.60	0.837	1.198	2.001	4.491	2.738	4.349
a_2 = 1.00	0.828	1.193	2.019	4.532	2.756	4.341
a_3 = 1.40	0.805	1.163	2.018	4.519	2.637	4.218
Item Difficulty						
(Low)-2.0to-0.8	0.781	1.148	2.050	4.682	2.612	4.184
(Med)-0.6to+0.6	0.845	1.202	1.998	4.443	2.782	4.371
(High)0.8to 2.0	0.845	1.197	1.994	4.416	2.772	4.353
Ability Distr.						
Equal	0.826	1.187	2.009	4.500	2.730	4.312
Unequal	0.820	1.178	2.019	4.527	2.716	4.293

At low sample size, high discrimination, and low item difficulty the cutoffs, would yield more false positive identification of biased items than the use of tabled values. The rest of the levels in the independent variables have mean P_{90} above the tabled value.

For P_{95}, the values were above the (tabled) expected values of 3.84. This implies that using the tabled value for P_{95} would result in more items being identified as biased .

Effect of the Independent Variable on the Distribution of MH-delta.

Following a significant MANOVA effect, univariate ANOVA was conducted on the mean of MH-delta. Two significant main effects of sample size and item difficulty occurred for the means of MH-delta. A multivariate and univariate significant four-way interaction of sample size by ability distribution by item difficulty by discrimination occurred.

Since the variables showing significant main effects were also included in the four way interaction, only the four-way interaction was analyzed. In order to analyze the overall effects, four one-way ANOVAs were conducted for each of the independent variables contributing to the interaction effect.

After significant one-way ANOVAs, Scheffe's test of pairwise comparisons was conducted for each independent variable. At item difficulty levels, group means of MH-delta for low item difficulty and high item difficulty were significantly different. So were the low and medium levels of item difficulty.The mean at medium item difficulty was the

lowest,

Table 8

Descriptive Statistics of MH-delta over 100 Replications across the Independent Variables.

	Mean	Std	Skewn	Kurt	P_{90}	P_{95}
Sample Size						
300/100	0.019	0.845	0.478	0.915	1.449	1.955
600/200	0.008	0.553	0.429	0.797	0.966	1.229
1200/400	0.003	0.377	0.415	0.767	0.657	0.823
Discrimination						
a_1 = 0.60	0.005	0.501	0.400	0.786	0.870	1.091
a_2 = 1.00	0.010	0.591	0.440	0.823	1.023	1.326
a_3 = 1.40	0.015	0.648	0.470	0.872	1.180	1.590
Item Difficulty						
(Low)-2.0to-0.8	0.040	0.773	0.516	0.989	1.330	1.830
(Med)-0.6to+0.6	0.002	0.501	0.403	0.747	0.875	1.094
(High)0.8to 2.0	0.011	0.502	0.404	0.744	0.867	1.084
Ability Distr.						
Equal	0.012	0.595	0.431	0.836	1.027	1.346
Unequal	0.008	0.589	0.442	0.817	1.021	1.326

while at low item difficulty the mean of MH-delta was the highest, (see Table 8).

For sample size, the group means of MH-delta were significantly different between small sample size of 400 and large sample size of 1600. Across discrimination and ability distribution levels there were no significant differences and based on the MANOVA results this is as expected.

For the standard deviation of the MH-delta a significant multivariate and univariate four-way interaction was found. There were three significant main effects of sample size, item difficulty, and discrimination. Because of the significant four-way interaction, four one-way ANOVAs were conducted, for each independent variable in the interaction. For each significant effect, the Scheffe test was conducted to determine which levels differed significantly. The means of the standard deviation of MH-delta for each level of each independent variable are shown on Table 8.

For sample size, all the pairwise comparisons using Scheffe test at 0.01 level of significance were significantly different. The standard deviation decreased with the increase in sample size. This implied that the index became more stable with sample size increase.

Similarly at the discrimination level of 0.6, 1.0, and 1.4, Scheffe's test for pairwise comparison were all significantly different. The standard deviation increased with the increase in discrimination level. At item difficulty the means of the standard deviation of MH-delta were significantly different between the low level and both medium and high levels of item difficulty. Between medium and high item difficulty there was no significant difference. No significant differences were found across ability

distribution levels.The four-way interaction may be interpreted cautiously as it is difficult and complex to interpret with certainty.

For the skewness values of MH-delta a multivariate and univariate significant three-way interaction of item difficulty by sample size by discrimination occurred. The three independent variables also showed significant main effect. Although significant two-way interaction also occurred, they were subsets of the three-way interaction and the analysis of the significant three-way interaction would include the significant two-way interaction that occurred.

So as to analyze the effects of the independent variables on the skewness one-way ANOVAs were conducted for each of the three independent variables in the interaction. Each one-way ANOVA significant effect was followed by a Scheffe test at 0.01 level of significance.

The item difficulty group mean of the skewness of MH-delta (see Table 8) at low item difficulty was significantly greater than that for both high item difficulty and medium item difficulty. However, there was no significant difference between medium and high item difficulty values. For discrimination the mean of skewness at high discrimination was significantly greater than at the low discrimination, while the remaining pairwise comparisons for discrimination were not significantly different.

In the case of sample size, the mean of skewness values at low sample size of 400, was significantly greater than for the large sample size 1600. Similarly the mean of the skewness values at low sample size was significantly greater than for the moderate sample size of 800. However, between the sample size of 800 and 1600, there was no

significant difference.

A two-way significant interaction between discrimination and ability distribution which was not a subset of the three-way interaction of sample size by item difficulty by discrimination occurred. For equal ability distribution, the mean at discrimination level of 1.4 was significantly greater than at the discrimination levels of 0.6 and 1.0. However, there was no significant difference between discrimination level of 0.6 and 1.0. (see Table 9).

Across unequal ability distribution there were no significant differences between discrimination levels. At all levels of discrimination, between the ability levels, there were no significant difference.

It can be deduced that the three variables, sample size, item difficulty, and discrimination have significant effects on the skewness of MH-delta. However, ability distribution was not a significant factor. Three-way interactions need to be interpreted with caution as they are difficulty to explain with certainty.

At the kurtosis values of MH-delta, a significant three way interaction effect of item difficulty by sample size by discrimination occurred. Sample size and item difficulty showed significant main effects. The other two-way interaction effects of item difficulty by sample size and item difficulty by discrimination were already included in the three-way interaction and will not be discussed.

Table 9

Group Means of Skewness of MH-delta at the Discrimination by Ability Distribution Interaction.

Ability Distribution	Discrimination Level			Group Difference
	0.6	1.0	1.4	
Equal Ability	0.3915	0.4331	0.4935	(1,3)(2,3)S
Unequal Ability	0.4285	0.4473	0.4501	N.S
Group Difference	N.S	N.S	N.S	

S =significant effect at 0.01 level

N.S = no significant effect at one-way ANOVA

Scheffe tests were conducted at 0.01 level significance.

So as to analyze the three-way interaction, three one-way ANOVAs were conducted, followed by Scheffe's test, for each of the independent variables in the interaction. For item difficulty, Scheffe tests on pairwise comparison at 0.01 level of significance showed that the mean of the kurtosis of MH-delta at low item difficulty and high item difficulty were significantly different(see Table 8). The kurtosis values were lowest at the high item difficulty and highest at low item difficulty; low item difficulty kurtosis values were significantly greater then medium item difficulty values. However, there was no significant difference between medium item difficulty and high item difficulty for kurtosis.

At sample size levels the low sample size of 400 had kurtosis values significantly greater than the moderate sample size and large sample size. However, there was no significant difference in the moderate sample size kurtosis values with those at high sample size (see Table 8). Scheffe tests also showed no significant differences on pairwise comparisons for discrimination.

Effect of the Independent Variables on P_{90} and P_{95} of MH-delta

A MANOVA was conducted followed by a univariate ANOVA on the P_{90} and P_{95} of the MH-delta, so as to analyze the effect of the independent variables on the false positive rate of 0.10 and 0.05 level.

For P_{90} of MH-delta, a three-way significant interaction effect of discrimination by item difficulty by sample size occurred. There were three two-way significant interaction effects of item difficulty by sample size and discrimination by difficulty. Since the interactions are subsets of the three-way interaction effect, none of them will be discussed. Three significant main effects also occurred, namely sample size, item difficulty and discrimination. Since these three variables are in the significant three-way interaction, their main effects will not be discussed here.

So as to analyze the three-way significant interaction effect of discrimination by item difficulty by sample size, one-way ANOVAs were conducted for each of the three independent variables. Significant effects were followed by Scheffe tests of pairwise comparisons at 0.01 level of significance. Results of the Scheffe test on pairwise comparisons for discrimination and sample size indicated that all the possible pairs of

comparisons were significantly different (see Table 10). In addition ,for item difficulty the mean for low item difficulty was significantly greater than for medium and high item difficulty.

Table 10

Group Means of MH-delta P_{90} for the three-way Interaction of Sample Size by Item Difficulty by Discrimination.

Variables	Low	Med.	High	Group Diff.
Discrimination	0.870	1.023	1.178	(1,3)(1,2)(2,3)S
Item Difficulty	1.330	0.875	0.867	(1,3)(1,2)S
Sample Size	1.440	0.966	0.657	(1,3)(1,2)(2,3)S

S = Significant effect at 0.01 level.

Scheffe tests were conducted at 0.01 level of significance.

It is evident from Table 10 that the three variables, sample size, item difficulty and discrimination, significantly affected the stability of the P_{90} of MH-delta. For discrimination level, the means of the percentiles increased from the low level of discrimination through the medium to the high level of discrimination.

At item difficulty levels, the mean P_{90} decreased from low item difficulty and was observed to be lowest at the high level of item difficulty. Across sample size the mean of the P_{90} decreased from the low sample size of 400 to the lowest at the large sample size of 1600.

For P_{95} of MH-delta, the MANOVA resulted in a multivariate and univariate

significant four-way interaction of all the four independent variables. Any other significant effects are subsets of the four way significant interaction effect. Three significant main effects also occurred. These were, sample size, item difficulty and discrimination. Ability distribution was not a significant main effect. However, ability distribution affected the P_{95} of MH-delta, through interaction with discrimination in a significant two way interaction effect as well as in the four-way interaction.

In order to analyze the four-way interaction effect four one-way ANOVAs were conducted followed by Scheffe tests for each independent variable in the four-way interaction. Table 11 shows the means of the P_{95} of MH-delta at each level for each of the four variables.

Table 11

Group Means of P_{95} of MH-delta for the four-way Interaction of Sample Size by Discrimination by Item Difficulty by Ability Distribution.

Independent Variable	Level 1 Low	Level 2 Med	Level 3 High	Group Difference
Discrimination	1.091	1.326	1.590	(1,2)(1,3)(2,3)S
Item Difficulty	1.830	1.094	1.084	(1,2)(1,3)S
Sample Size	1.955	1.229	0.823	(1,2)(1,3)S
Ability Distribution Levels	1.346	1.336	-----	N.S

S =significant effect at 0.01 level.

N.S = no significant effect for one-way ANOVA

Scheffe tests were conducted at 0.01 level of significance.

Significant pairwise differences for each independent variable are shown on Table 11. At the discrimination levels the mean P_{95} at high discrimination was greater than for medium and low levels of discrimination. All the pairs were significantly different with values increasing from low to high level of discrimination. For both item difficulty and sample size, Scheffe's test of pairwise comparisons of the levels, shows that the mean of P_{95} at the low level were significantly greater than the mean of P_{95} at medium and high level. There were no significant differences between the means at medium and high levels for both item difficulty and sample size.

While it can be concluded that the means of P_{95} of MH-delta were significantly influenced by the four independent variables jointly, the interpretation of the four-way interaction poses some difficulties. Sample size and item difficulty showed significant main effects in the distribution of the index and in P_{90} and P_{95} . However, discrimination which showed a significant main effect in the distribution of MH-delta influenced the percentiles through its interaction effect with ability distribution. The false positive rates corresponding to the percentiles can be deduced to be affected in the same way.

The independent variables affected the distribution of MH-delta and the stated percentiles in a consistent manner, although ability distribution had no unique effects on the distribution and the percentiles. The three-way and four-way interactions posed some

difficulties and limitations to the interpretation of their results. However, this was resolved by conducting a separate one-way ANOVA for each interaction variable.

The results of this study agree with Gutierrez (1989) in which sample size, discrimination, and item difficulty interacted significantly, affecting the stability of the MH-delta index. Sample size was found to be consistently affecting both the distribution and the percentiles of MH-delta.

Since the distribution of MH-delta is not known the stability of the index under the effects of the independent variable is best analyzed through the mean, standard deviation, P_{90}, and P_{95}. This is because, when baseline studies are conducted, cutoffs are established at the stated percentiles or units of standard deviations in the identification and analysis of biased items. Examination of skewness and kurtosis may provide further information as to the nature of the distribution of MH-delta.

In instances where no known associated statistical tests exist and the distribution of the index is unknown, cutoffs in terms of percentiles and standard deviations provide the only alternative in dif analysis. Table 8 provides a summary of the descriptive statistics of MH-delta. The means are close to the expected value of zero. The standard deviation, skewness, and kurtosis are not known and so the computed values can not be compared to any expected values. Under these conditions the practitioner would be compelled to use baseline studies to set cutoffs for specific false positive rates.

Effects of the Independent Variables on the Distribution of LU.

The MANOVA results on the descriptive statistics of the LU index showed no

Table 12

Descriptive Statistics of LU across the Independent Variables over 100 Replications

	Mean	Std	Skewn	Kurt	P_{90}	P_{95}
Sample Size						
300/100	1.001	1.334	1.929	3.263	3.149	4.617
600/200	1.006	1.326	1.880	3.075	3.148	4.560
1200/400	0.999	1.317	1.895	3.170	3.127	4.592
Discrimination						
a_1 = 0.60	0.996	1.322	1.919	3.224	3.103	4.583
a_2 = 1.00	1.011	1.341	1.910	3.201	3.193	4.633
a_3 = 1.40	0.998	1.315	1.874	3.082	3.129	4.550
Item Difficulty						
(Low)-2.0to-0.8	1.004	1.334	1.920	3.230	3.133	4.586
(Med)-0.6to+0.6	1.007	1.336	1.897	3.123	3.178	4.607
(High)0.8to 2.0	0.994	1.308	1.886	3.145	3.115	4.576
Ability Distr.						
Equal	1.007	1.336	1.903	3.168	3.149	4.636
Unequal	0.996	1.316	1.900	3.170	3.134	4.543

significant effects of the independent variables on the mean, standard deviation, skewness, and kurtosis of the LU index. The means of LU however were close to the expected value of 1.0 (see Table 12) the expected mean of a chi-square distribution with

one degree of freedom. The standard deviation was slightly underestimated and fell below the expected value of 1.40. Skewness values were substantially below the expected value of 2.82.

Effect of the Independent Variables on P_{90} and P_{95} of LU.

As with the distribution of LU, no significant effects occurred at both the P_{90} and P_{95}. Table 12 shows the means of the P_{90} and P_{95} at each level of the four independent variables. False positive rates at 0.1 and 0.05 levels for LU were therefore not significantly affected by sample size, item difficulty, discrimination, and ability distribution. For P_{90}, the mean values of the percentile are larger than the expected (Tabled) value of 2.70, at 0.10 level of significance. For all the independent variables, they ranged from 3.075 to 3.263. This implies that the use of tabled values would result in more false positive identification, compared to the values computed in this study. Similarly, the means of the P_{95} were larger than the expected value of 3.84 at 0.05 level of significance, for all levels of the independent variables. The same result would occur if tabled values were used in the detection of biased items; that is, more false positives would be identified by the use of tabled value compared to the values obtained in Table 12.If the observed data reflected reality then a baseline study would be appropriate using more than one replication as shown in the literature to establish where the cutoff will be.

The distribution of the LU index and the two percentiles studied were not significantly affected by sample size, discrimination, item difficulty, or ability distribution. LU is therefore robust to the effects of the stated independent variables and is a very

stable index. However it was observed that the standard deviation, skewness, and kurtosis values were notably underestimated while P_{90} and P_{95} were overestimated relative to the expected values. A possible explanation could be an effect of the estimation program used (Spray,1989b).

Power studies conducted by Rogers and Swaminathan (1990) showed that LU was significantly affected by model data fit and test length. These two variables were not studied here. Although sample size was found to affect the detection rate of LU in the Rogers and Swaminathan study (1990), the findings in this study show that sample size did not significantly affect the distribution of LU index and the percentiles (P_{90} and P_{95}).

Effect of the Independent Variables on the Distribution of LN.

For the means of the LN index, a multivariate and unvariate significant two-way interaction effect of discrimination by item difficulty occurred. There were three significant main effects of sample size, item difficulty and discrimination. Since item difficulty and discrimination occurred in the two-way interaction their main effects will not be discussed. However sample size main effects were analyzed. The Scheffe test was conducted so as to determine which levels of sample size differed significantly. A significant difference between means at sample size of 400 ($m=1.085$) and sample size of 1600 ($m=1.033$) was found.

Table 13

Descriptive Statistics of LN across the Independent Variables over 100 Replications.

	Mean	Std	Skewn	Kurt	P_{90}	P_{95}
Sample Size						
300/100	1.085	1.451	1.920	3.190	3.423	4.921
600/200	1.051	1.409	1.901	3.215	3.333	4.796
1200/400	1.033	1.390	1.913	3.124	3.264	4.740
Discrimination						
a_1 = 0.60	1.012	1.353	1.894	3.121	3.196	4.668
a_2 = 1.00	1.050	1.402	1.907	3.131	3.301	4.777
a_3 = 1.40	1.115	1.495	1.933	3.277	3.523	5.012
Item Difficulty						
(Low)-2.0to-0.8	1.030	1.382	1.908	3.151	3.254	4.200
(Med)-0.6to+0.6	1.014	1.364	1.923	3.200	3.203	4.729
(High)0.8to 2.0	1.131	1.500	1.904	3.179	3.563	5.031
Ability Distr.						
Equal	1.052	1.419	1.915	3.132	3.340	4.820
Unequal	1.059	1.415	1.908	3.221	3.339	4.810

The mean of LN at low sample size of 400 through the medium sample size of 800 to the large sample size of 1600, decreased with the increase in sample size, although the

only significant increase was as stated (between low and large sample size). Sample size is therefore a significant factor that affected the LN index.

The means of LN were close to the expected value of 1.0, the value of the mean of a chi-square distribution with one degree of freedom(see Table 13).

The significant two-way interaction of item difficulty by discrimination that occurred for the mean of the LN index was analyzed using simple effects, followed by Scheffe's test. Table 14 shows the group means of LN for the discrimination by item difficulty interaction.

Table 14

Group Means of LN for the Discrimination by Item Difficulty Interaction

Discrimination Level	Item Difficulty Level			Group Difference
	Low	Med.	High	
a_1 = 0.6	0.999	0.997	1.039	N.S
a_2 = 1.0	1.023	1.006	1.117	(1,3)(2,3)S
a_3 = 1.4	1.066	1.040	1.233	(1,3)(2,3)S
Group Difference	(1,3)S	N.S	(1,2)(1,3)(2,3) S	

N.S = not significant one way ANOVA

S = significant at 0.01 level

Scheffe tests were conducted at 0.01 level of significance.

Across low item difficulty level, the group mean at discrimination level of 1.4 was significantly greater than at discrimination level of 0.6. For medium item difficulty the pairwise comparisons of group means at the discrimination levels of LN were not significantly different. However at high item difficulty all the pairwise comparisons were significantly different. For high item difficulty the mean values of LN were all larger than the expected value of 1.0. A possible explanation is that more guessing could have taken place especially at the steeper slope (discrimination of 1.4) with more low ability candidates guessing randomly.

Across the discrimination level of 0.6, the group means for item difficulty were not significantly different. However, as shown in Table 14, across 1.0 and 1.4 the pairwise comparisons of the means at each of these levels indicated that high item difficulty means were significantly greater than for medium and low item difficulty.

For the standard deviation of LN, as in the case of the means, there were significant main effects of sample size, item difficulty and discrimination. A multivariate and univariate significant two-way interaction of discrimination by item difficulty also occurred for the standard deviation of the LN index.

Sample size as a significant main effect of the standard deviation of LN, was analyzed using Scheffe's test. The pairwise comparisons showed that the group mean of standard deviation at low sample size (m=1.451) was significantly greater than the standard deviation group mean at high sample size (m=1.390). As the sample size increased the standard deviation means decreased, showing that the LN index became more stable (see Table 13).

Since discrimination and item difficulty occurred as a significant two-way interaction, their main effects will not be discussed. Instead, the two-way interaction effects were analyzed using simple effects analysis (see Table 15).

Table 15

Group Means of the Standard Deviation of LN for the two-way Interaction of Item Difficulty by Discrimination.

Discrimination Level	Item Difficulty Level			Group Difference
	Low	Med.	High	
a_1 = 0.6	1.341	1.324	1.394	N.S
a_2 = 1.0	1.368	1.370	1.466	(1,3)(2,3)S
a_3 = 1.4	1.437	1.390	1.650	(1,3)(2,3)S
Group Difference	(1,3)S	N.S	(1,3) (2,3)S	

S = significant at 0.01 level

N.S = not significant one-way

Scheffe test conducted at 0.01 level of significance.

At a discrimination level of 0.6 the group means of the standard deviation of LN across item difficulty were not significantly different. At discrimination levels of 1.0 and 1.4, there were significant pairwise differences for low and medium item difficulty as

compared to high item difficulty which had significantly larger values(see Table 15).

From Table 15 it is observed that at low item difficulty across all the discrimination levels, the Scheffe test of pairwise comparisons resulted in a significant difference between discrimination level of 0.6 and 1.4 . At medium item difficulty, across discrimination level there was no significant difference. At high item difficulty, the Scheffe test showed that values at discrimination level of 0.6 and 1.0 were both significantly lower than at 1.4.

The standard deviation values of LN (see Table 13) were close to the expected value of 1.4, which is the standard deviation for a chi-square distribution with one degree of freedom. At high item difficulty and high discrimination, the standard deviation values were higher than expected value. This implied that the index was unstable at the high item difficulty and high discrimination as it showed relatively high variability.

It can be concluded that discrimination, item difficulty, and sample size have significant effects on the standard deviation of LN. For skewness and kurtosis of LN, there were no significant effects for the independent variables.

Effects of the Independent Variables on the P_{90} and P_{95} of LN.

At P_{90} of the LN index, sample size showed a significant main effect. So did item difficulty and discrimination. A multivariate and univariate significant two-way interaction of discrimination by item difficulty also occurred at P_{90}.

So as to analyze the main effects of sample size, Scheffe tests following the univariate result were conducted on the means of P_{90}. Low sample size of 400 (m=3.423)

had a mean of P_{90} that differed significantly with those at large sample size of 1600 (m=3.264). The means of the percentiles decreased with an increase in sample size.However at all sample size levels means of P_{90} were higher than the expected value of 2.70 at 0.10 level of significance. The use of the tabled value would result in more false positives than the use of the computed values in this study.

Table 16

Group Means of P_{90} of LN for the Discrimination by Item Difficulty two-way Interaction

Discrimination Level	Item Difficulty Level			Group Difference
	Low	Med.	High	
a_1 = 0.6	3.215	3.122	3.250	N.S
a_2 = 1.0	3.225	3.192	3.490	(1,3)(2,3)S
a_3 = 1.4	3.320	3.297	3.952	(1,3)(2,3)S
Group Difference	N.S	N.S	(1,3) (2,3)S	

S = significant at 0.01 level

N.S = not significant one-way ANOVA

Scheffe test conducted at 0.01 level of significance.

The significant two-way interaction effects of discrimination by item difficulty was analysed by conducting simple effects analysis followed by the Scheffe test. Table 16 shows the group means of P_{90} of the LN index for the discrimination by item difficulty interaction.

At low and medium item difficulty there were no significant differences found for discrimination. However group means at high item difficulty, across discrimination level 0.6 and 1.0 were significantly lower than at discrimination level of 1.4. At 1.0 and 1.4 discrimination levels, P_{90} at both low and medium item difficulty was significantly lower than for high item difficulty.

It can be concluded that the three independent variables, (sample size,item difficulty,and item discrimination) significantly affected P_{90} for LN. Sample size as a significant main effect, and item difficulty and item discrimination through the interaction, significantly affected P_{90} of LN.

At P_{95} of the LN index, there were no significant interaction effects among the independent variables. However, two significant main effects of discrimination and item difficulty occurred.In order to analyze the effect of discrimination on P_{95} of the LN index, Scheffe's test was conducted to determine which level of discrimination differed significantly. There were significant differences at discrimination level between 0.6 and 1.4 and between 1.0 and 1.4 levels. The group means of the P_{95} at discrimination of 0.6 was 4.67 while at discrimination of 1.0, the mean was 4.78. At the high discrimination level the mean P_{95} was 5.012.

In order to analyse the significant effects of item difficulty Scheffe's test was

conducted. The pairwise comparisons showed that the group mean of P_{95} for LN at high item difficulty was significantly greater than at both medium and low item difficulty. The mean P_{95} of LN for all the independent variables was larger than the tabled value of 3.84 which is the expected cutoff value at the 0.05 level of significance for a chi-square distribution with one degree of freedom. The use of the tabled value of 3.84 would result in more false positives than when the computed values at P_{95} of LN.The stability of the LN index which is inferred from the standard deviation was significantly affected by sample size and by interaction of item difficulty with discrimination.

In the case of P_{90}, sample size, and the two-way interaction effects of item, difficulty by discrimination were found to show significant effects. However, at P_{95} of LN sample size was not significant but discrimination and item difficulty showed significant main effects.

The interaction effects of discrimination and item difficulty were found to affect the distribution through the effects on the mean and standard deviation of the LN index, and the P_{90} in the same way. The expected (tabled) values of the P_{90} and the P_{95} of LN of 2.70 and 3.84, respectively, were lower than the values computed and obtained in this study (see Table 13). This has an important implication in that tabled values would have a higher false positive identification than P_{90} and P_{95}.

The significant effect of sample size on the distribution of LN as well as the P_{90} of LN is in agreement with Swaminathan and Rogers (1990) findings, in that with the increase in sample size, the proportion of false positive identification decreased. It was also further noted that with increase in sample size the LN standard deviation reduced

and closely approached the expected value of 1.4. This result indicated that the LN index became more stable at high sample size.In the study by Swaminathan and Rogers (1990),test length was found to significantly affect LN in that longer tests had fewer false positives than shorter tests.In the present study test length was not studied as an independent variable.

The summary and conclusions of this study are presented in the next chapter.

CHAPTER V

SUMMARY AND CONCLUSIONS

The results of this study show that the LU index and the LN index have advantages over the Mantel-Haenszel procedures (MH-delta and MH-CHISQ). The first advantage is the fact that sample size, item difficulty, discrimination, and differences in ability distribution for reference and focal groups did not have significant effects on LU.This is an important quality of a reliable and stable index that would be most suitable in DIF detection and analysis.

Although Rogers and Swaminathan (1990) and Swaminathan and Rogers (1990) showed that MH-CHISQ and LU were equally efficient in detecting uniform DIF, the fact that LU is not significantly affected by the stated independent variables is an advantage over MH-CHISQ and certainly over MH-delta.

Sample size showed significant main effects for all the dependent variables except for LU and at kurtosis as well as skewness for LN. Sample size was the most consistently significant independent variable for the distribution and percentiles (P_{90} and P_{95}) of the indices. Ability distribution showed no significant main effect for any of the four indices studied.

For the MH-CHISQ index, item discrimination showed a significant main effect only for the means of the MH-CHISQ index but not at any of the two percentiles, standard deviation, skewness and kurtosis of MH-CHISQ. For MH-delta, item

discrimination showed significant main effects for P_{90}, P_{95}, the standard deviation, and skewness of MH-delta. However, the variable did not show significant effects for the mean and kurtosis of MH-delta. For LU, discrimination had no significant effect for any of the dependent variables. For LN, discrimination showed significant main effects at both P_{90} and P_{95}, the mean, and standard deviation.However discrimination showed no significant effects for kurtosis and skewness of LN index.

Item difficulty showed significant main effects for P_{90}, mean, and skewness of MH-CHISQ. For the MH-delta index, item difficulty showed a significant main effect for all the six descriptive statistics used as the dependent variables. However, item difficulty was not significant for any of the LU dependent variables. For LN, item difficulty showed significant main effects at P_{90}, P_{95}, the mean, and standard deviation. No significant effects were found for kurtosis and skewness of LN.

For large sample size the P_{90} and P_{95} values of MH-CHISQ exceeded the tabled values. At low sample size and high discrimination, P_{90} of MH-CHISQ was lower than the expected (tabled) value of 2.70.In this case the use of the computed values of P_{90} of the MH-CHISQ would result in more false positives than when the tabled value is used. However, at any other level of the independent variables, P_{90} of MH-CHISQ was overestimated and larger than the tabled values of 2.70 . P_{95} of MH-CHISQ was also overestimated for all the levels of the independent variables and was larger than the tabled value of 3.84.

For MH-delta, the means were close to the expected value of zero. Since the distribution of MH-delta is not known, and P_{90} and P_{95} varied across the independent

variables, the use of the percentiles (P_{90} and P_{95}) should be based on baseline studies using several replications for stable results upon which cutoffs would be established.

For LU, the means were close to the expected value of one, while the standard deviations were slightly lower than the expected value of 1.4. P_{90} and P_{95} were larger than expected (tabled)values of 2.70 and 3.84 respectively. This implied that the use of tabled values would result in more false positives than the use of the computed P_{90} and P_{95} cutoffs in these data.

For LN, the means were close to the expected value of one, throughout the levels of independent variables. The standard deviations were also close to the expected value of 1.4 except at high discrimination and high item difficulty where the standard deviation was overestimated. This implies that the index was unstable at these levels of the independent variables. P_{90} of LN was larger than the tabled value of 2.7. The P_{90} ranged from 3.196 to 3.563, with the lowest value being found at the low discrimination level and the highest at high item difficulty. As for P_{95}, the values were once again higher than the tabled value of 3.84, with the lowest value at the low discrimination (m=4.67) and the highest value occurring at high item difficulty (m=5.031). Like in the case of LU, the LN cutoffs at P_{90} and P_{95} were larger than the expected values and so the use of tabled or expected values would result in more false positives than when the computed values are used.

In conclusion, the LU and LN indices of the logistic regression procedure are the most appropriate to use in DIF detection and analysis, taking into account, the stability of the indices and the fact that the LU index is robust to variations in the stated

independent variables. In addition to this, the LU and LN indices have known distributions unlike MH-delta. The means of LU and LN were very close to the expected values although the standard deviations of LU and LN were underestimated in some cases. The skewness values and the kurtosis are in agreement with the expectations of a chi-square distribution, although they were also underestimated. The kurtosis in both indices was both positive, resulting in leptokurtic distributions, for LN and LU. However, there were significant differences between observed and expected distributions of the MH-CHISQ, LU, and LN indices. A possible explanation of the under estimation could be attributed to the estimation program (Spray 1989b).This should be investigated.

The cutoffs obtained at P_{90} and P_{95} for LU and LN would be of interest to practitioners as their use would result in lower false positive rates than the tabled values.The fact that the two percentiles of LN were significantly affected by the independent variables, implies that the LR procedure should be used with caution.

Relationship of Findings to other Research.

Simulation studies on the distribution of MH indices and LR indices are few and limited. Even fewer are comparative studies of the two procedures of LR and MH. Guiterrez (1989) conducted a Monte Carlo study to examine the effects of sample size, item difficulty, and item discrimination on the distribution of MH-delta.
The findings of the study were, that sample size, item discrimination, and item difficulty had significant effects on the standard deviation of MH-delta. It was further noted that for all combinations of sample size, and values of discrimination studied, the standard

deviation of MH-delta was significantly larger when item difficulty values were located at the extremes(low and high item difficulty) than at medium level of difficulty. The findings showed that MH-delta is unstable at the extreme levels of item difficulty. This is in agreement with this study. In this study unstable standard deviations, P_{90}, and P_{95}, of MH-delta were observed at low item difficulty and at high levels of item difficulty. In both Guiterrez study and this study it was found that sample size and item discrimination had significant effects on the distribution and the percentiles of MH-delta.

Rogers and Swaminathan (1990) carried out a simulation study to examine the distributional properties of the LR and MH test statistics and to investigate the relative power of LR and MH. Their findings showed that MH and LR did not have the expected distribution at high item difficulty and high item discrimination level, as indicated by Kolmogorov-Smirnov test. The difference between expected and observed distribution for the two test statistics was attributed to model data fit and the independent variables. The study also examined the dif detection rate of both MH and LR. Results showed that the LR procedure is more powerful than MH for detecting nonuniform dif.

The results of the study are in partial agreement with this study in terms of the effects of item difficulty and discrimination on the distribution of both LR and MH. However, the reasons for differences in observed distribution and expected distribution of the two procedures differ.In the case of Rogers and Swaminathan (1990) study, model data fit as one of the independent variables significantly affected the distributions, while in the case of this study the stated independent variables were found to significantly

affect the distributions of the indices.This study did not compare the power of the indices since dif was not induced in the data.

In conclusion, this study found that for MH-CHISQ, LU, and LN the P_{90}, and P_{95} were larger than the tabled values across the independent variables. These findings have significant educational implications.Sample size and the three other independent variables studied had no significant effect on LU. However, the P_{90} and P_{95} of LU were lager than tabled values.

Ability distribution was found to have no significant effect on the indices except as a moderator across the discrimination variable for MH-CHISQ. This result differs from Mazor et.al.(1991) in which it was stated that ability distribution had significant effects on the MH statistic. The other significant finding of this study is that the three indices LN, MH-CHISQ, and MH-delta were significantly affected by sample size, item difficulty, and item discrimination. The observed and expected distribution of LN, LU and MH-CHISQ were found to be different.

It was also found that the means of MH-delta and LN were overestimated while the mean and standard deviation of MH-CHISQ were underestimated. In the case of LU the standard deviation was underestimated. Similarly, P_{95} and P_{90} of MH-CHISQ, LU, and LN were overestimated although, for low sample size and high discrimination the P_{90} of MH-CHISQ were notably lower than the tabled values. The P_{90} and P_{95} of MH-delta also varied across the independent variables.

Limitations and Future Research

The results of this study were based on simulated data under the null hypothesis. It would be realistic to the practitioner if the study were conducted with data in which known dif is induced.Real data can not be used due to the fact that true state of dif is not known. Another limitation is the fact that only four independent variables, namely sample size, discrimination, item difficulty and ability distribution, were studied with selected and restricted ranges of values. This implies that the results obtained may be specific to these data and conditions of the independent variables stated. In the case of real data from actual test administration more variables other than those stated here such as test length, contribute to dif. A replication of this study would be appropriate if the findings are to be generalized to real data. Studies by Rogers and Swaminathan (1990) showed that test length had an effect on LU and LN. In the present study test length was not studied as an independent variable.

Further research needs to be conducted incorporating other variables such as test length, with the four independent variables. The use of different values, other variables, and ranges of the four independent varibles should also be researched. Although this study has shown that LU and LN indices are the most appropriate to use, the use of these indices among practitioners is limited due to the cost and availability of suitable, yet simple computer programmes. Since the estimation programme used (Spray 1989b) is suspected to underestimate certain values and overestimate some, the program should be researched in an effort to improve its performance and accuracy. Finally, research needs to be done in which dif is induced into the data so as to analyse the

correct detection rate for each of the four indices, in order to compare and determine which of the indices is most efficient and stable across the variables studied.

Educational Implications.

The MH and LR procedures have advantages and disadvantages .The advantage of MH is its computational simplicity and relatively low cost. This has made it common and popular among test developers. However, MH has a limitation in that although it detects uniform dif, it fails where there is interaction between ability (the matching criterion) and group membership. For this reason MH is a weak detector of nonuniform dif.The performance and distribution of MH indices were observed to be significantly affected by sample size, item difficulty, and item discrimination.

The LR procedure has the advantage of detecting both uniform and nonuniform dif, as well as being relatively less affected by the independent variables stated (LU in particular). However, the iterative and complex nature of its computation makes LR relatively more expensive in terms of computer time and software. A significant point to note is that experts of measurement and educators alike should be aware of the limitations of MH and LR procedures as shown in this study.

In the case of MH-delta, little is known of the distribution. This limits the index to use as an indicator of the direction of bias. The relationship of the magnitude of MH-delta to the magnitude of bias in an item is not clear, since the distribution of the index is unknown. Therefore, no logical deduction can be made about the magnitude of MH-delta as an indicator of the magnitude of bias in an item. Moreover the MH procedure

in general is designed to detect only uniform bias. This implies that nonuniform bias may not be detected well by MH procedures. The LR procedure detects both uniform and nonuniform bias and therefore has an advantage over the MH procedure. In addition to this, the four independent variables studied were found to have significant effects on the performance of MH-delta and MH-CHISQ but relatively less effect on LU and LN.

The effects of the independent variables (sample size, item difficulty, item discrimination, and ability distribution) should be considered whenever one intends to use the MH and LR procedures and correctly interpret the results. The effects of the independent variables on the P_{90} and P_{95} of the indices has resulted in larger percentiles compared to the tabled values. This implies that the use of the tabled values would result in more items being labeled as false positives. This problem can be overcome by carrying out baseline studies using several replications or samples.

When analysing dif using MH-delta, ETS(Educational Test services) sets standards at specific values (Zwick & Ercikan,1989), irrespective of the effects of the independent variables on MH-delta. This poses a threat to the validity of the interpretetion of the dif identification conducted. The effects of these variables need to be taken into account whenever the MH procedure is being used.

The ACT (American College Testing) examination board has adopted LR procedure for dif analysis. However, any user intending to utilize LR should address himself or herself to the following concerns. First, the effect of the independent variables found in this study. Secondly, the high P_{90}, and P_{95} values for LU and LN that were noted to be larger than the tabled values, and thirdly the possiblity of the estimation

programme consistently underestimating some descriptive statistics.

Although it can be suggested that LR is a better procedure than MH, the former should be used with caution as evidenced by findings of this study. Whenever MH-Z is used in dif analysis, a baseline study should be conducted for a sample of the subpopulations involved. Such baseline studies should be conducted using more than one replication so as to improve the stability of the cutoffs. Since LR is least affected by the independent variables as compared to MH, it is recommended that LR be adopted by educators and measurement experts for dif analysis.

References

Ackerman, T. (1987). Program MANTEL, revised version.

Camilli, G., & Smith, J.K. (1990). Comparison of the Mantel-Haenszel Test with Randomised and Jackknife Tests for detecting bias in items. Journal of Educational Statistics, 15, 53-67.

Carlson, J. (1983). Program DATAGEN. IBM version of DATAGEN modified by J.Carlson.

Cleary, T.A. (1968). Test bias; Prediction of grades of negro and white students in integrated Colleges. Journal of Educational Measurement, 5, 115-125.

Crocker, L., & Algina, J. (1986). Introduction to Classical and Modern Test Theory. Holt, Rinehart and Winston.Orlando, Florida.

Donoghue, J.R., & Allen, N.L. (1991, April). 'Thin' versus 'Thick' matching in the Mantel Haenszel procedure for detecting dif: A Monte Carlo Study. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Gutierrez, J. (1989). Characteristics of the distribution of the Mantel-Haenszel delta under different conditions of the null hypothesis; A Monte Carlo Study. Thesis (MA), unpublished. University of Ottawa.

Hambleton, R.K., & Rogers, J.H. (1989). Evaluation of Computer Simulated baseline statistics for use in item bias studies.Educational and Psychological measurement , 49.

Hambleton, R.K., & Rovinelli.R. (1973). DATAGEN program.A Fotran iv program for generating examinee response data from logistic test models.Behavioral Science, 18, 73-74.

Hambleton, R.K., & Swaminathan, H. (1985). Item Response Theory; Principles and Applications. Nijhoff Publishing Company . Boston.

Holland, P.W(1985, september).The study of the differential item performance without IRT. Proceedings of the military Testing Association. San Diego, CA.

Holland, P.W., & Thayer, D.T. (1986). Differential item functioning and the Mantel-Haenszel procedure. Program statistics Research Technical Report no. 86-69, Research report 86-31. Princeton, NJ: Educational Testing Services.

Mazor, K.M., Clauser. B.S., & Hambleton, R.K. (1991, April). Examination of various influences on the Mantel-Haenszel statistics. Paper presented at the annual meeting of the American Educational Research Association. Chicago, ED 331876 (ERIC).

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute , 22, 719-748.

McPeek, W., & Wild, C.L(1986, April).Performance of the Mantel-Haenszel statistics in a variety of situations.Paper presented at the annual meeting of the American Educational Research Association. San Franscisco, CA .

Mellenberg, G.J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Phillips, S.E., & Mehrens, W.A. (1988, April). Comparison of methods for detecting differential item performance due to instructional/test misalignment. Paper presented at the American Educational Research Association annual meeting, New Orleans.

Raju, N.S., Bode, R.U., & Larsen, V.S. (1989): An empirical assessment of Mantel-Haenszel statistics for studying differential item performance. Applied measurement in Education, 2, 1-13.

Rogers, J., & Swaminathan, H. (1990, April).A comparison of Logistic Regression and Mantel-Haenszel procedures for detecting item functioning.Paper presented at the annual meeting of American Educational Research Association, Boston, M.A.

Spray, J.A. (1989, October).Performance of the three conditional dif statistics in detecting differential item functioning on simulated tests.ACT Research Report Series no.89-7.

Spray, J.A .(1989b).Estimation program of the logistic regression dif statistic.

Swaminathan,H., & Rogers, J.H (1990).Detecting differential item functioning using Logistic Regression procedure. Journal of Eductional Measurement,27 (4),361-370.

Sykes, R.C., & Fitzpatrick, A.R. (1990, April). Establishing a Mantel-Haenszel alpha cutscore through a multiple-method procedure. Paper presented at the annual meeting AERA, Boston, MA.

Zwick, R., & Ercikan, K.(1989): Analysis of differential item functioning in the NAEP history assessment.Journal of Educational Measurement, 26,(1), 55-66.