National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

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

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

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

AVIS

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

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

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

Canada

THE PERFORMANCE OF THE MANTEL-HAENSZEL AND LOGISTIC REGRESSION DIF IDENTIFICATION PROCEDURES WITH REAL DATA

By

Fang Tian

Thesis presented to the School of Graduate
Studies and Research as partial fulfilment
of the M.A. degree in Education

Faculty of Education, University of Ottawa
Ottawa, Canada, 1994

ISBN 0-612-00623-9

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ACKNOWLEDGMENTS

This thesis was completed under the supervision of Dr. Marvin W. Boss. His availability, suggestions, patience, and endless encouragement were greatly appreciated throughout the past two years. I would like to take this opportunity to thank him for being there not only as an advisor but also as a friend. I have learned a great deal from him.

I would also like to thank Dr. Marc E. Gessaroli and Dr. Bruno Zumbo for their helpful criticisms and suggestions. Their comments were very useful and contributed a great deal to this thesis.

I am also grateful to Mr. Michel Brabant for his valuable computer help.

Finally, a very special thank you goes to Ying for his support and patience in the completion of this work.

TABLE OF CONTENTS

Page

LIST OF TABLES

ABSTRACT

Numerous statistical methods have been proposed for detecting differential item functioning (DIF). Among them, methods based on item response theory (IRT) are theoretically preferred but very complicated and expensive to implement. As an alternative, the Mantel-Haenszel (MH) procedure has emerged as one of the most popular procedures because of its ease of implementation, relatively small sample size requirement, and associated test of significance. In addition, it provides a measure of the amount and direction of DIF. However, the MH procedure is not designed for and therefore not very effective in detecting nonuniform DIF. As an extension of the MH procedure, a more general DIF detection method, a logistic regression procedure (LR) has been shown to be powerful in detecting both uniform and nonuniform DIF.

The purpose of this study is to examine the consistency of the MH and LR procedures and their agreement in the identification of DIF across sample size and criterion when using real examinee data. The item responses of 183,008 Caucasian examinees who took the 1989 ACT Reading Test (Form 39B) were analyzed. The focus of this study was on gender-based DIF. Comparison groups were created by randomly selecting cases by gender across four levels of sample size with an equal number of examinees in both reference and focal groups: 100/100, 250/250, 500/500, and 1000/1000. Three criteria were used: total test score, subtest score, and passage score. Thirty replications were conducted for each combination of sample size and criterion variable.

The mean and standard deviation of the MH-Z and frequency of DIF identification by the MH and LR chi-square statistics over 30 replications were computed for each item for each combination of sample size and criterion variable. The agreement between the MH and LR procedures and the consistency of the two procedures were examined according to the number of times an item was identified. The MH-Z was used to determine the direction and size of DIF. Items were classified into three categories based on the means of MH-Z over 30 replications at sample size of 1,000: 1. No Serious DIF, 2. Possible DIF, 3. Definite DIF. As for nonuniform DIF, items which had been detected at least 15 times and 20 times out of 30 replications by the LR nonuniform statistic at the .01 significance level were considered as Possible DIF and Definite DIF, respectively.

Using the total test score as the criterion, seven items were detected as showing uniform DIF (4 Definite DIF and 3 Possible DIF). When the criterion was subtest score, twelve items were identified (5 Definite DIF items and 7 Possible DIF items). Using the passage score as the criterion, only four items were identified (2 Definite DIF and 2 Possible DIF).

The influence of the criterion variable is very obvious. Different numbers of items and different items were identified when using different criteria. Only two items were identified under all three criteria. In addition, when the criterion variable was subtest score, the means of MH-Z were generally lower for items in Passages 1 and 2 and higher for items in Passages 3 and 4 than when the criterion was total test score. Similar results were

observed for the detection rates of the MH and LR Uniform chi-square statistics. When the criterion changed from total test to subtest, the frequency of DIF detection decreased for items in Passage 1 and 2 and increased for items in Passage 3 and 4. This is likely due to the fact that for each subtest females scored higher on one passage but lower on the other. The direction of DIF was such that it favoured the group scoring higher on the passage. When the criterion was passage score, an even larger decrease of mean MH-Z values occurred and the frequency of DIF identification was close to zero for most of the items. Using total and subtest scores as the criteria, DIF items identified in Passage 2 all favoured males. However, when passage score was used, the MH-Z picked up two items with opposite effect (favouring females) that were not detected for the total and subtests. Since items based on each passage are associated to a specific content area, there seemed to exist a passage or context effect. However, It was very hard to separate the passage effect from the test length effect since there are only ten items in each passage.

The multidimensionality of the criterion is likely the cause of the results of this study. The low correlations among the four reading passages seemed to suggest that different aspects of reading ability were being measured. Another possible reason might be test length. However, in some way, test length might not be the issue since there were more items identified when subtest was used as the criterion.

Sample size had a strong influence on the MH and LR procedures. As sample size increased, the power of the MH and LR statistics improved substantially. At small sample sizes (100 and 250 subjects per group), the detection rates were very low even when the effect size was large. The means of MH-Z were relatively consistent across sample size. However, the standard deviation over 30 replications of the MH-Z for each item increased as the sample size decreased.

The results of DIF detection by the LR nonuniform statistic indicated no items exhibited serious nonuniform DIF in this dataset.

A comparison of the frequency of DIF detection by the MH and LR Uniform procedures shows that in general, the agreement between the two procedures is very high in the detection of uniform DIF with the LR procedure having a slightly advantage. As sample size decreased, the discrepancy between the two procedures became larger.

In conclusion, the results of this study supported the use of both the MH and LR procedures as practical means of detecting DIF. It was recommended that a sample size between 500 and 1,000 be used in order to obtain reliable MH and LR statistics. The results also indicated that there is substantial agreement between the MH and LR procedures in the detection of uniform DIF. Finally, this study provided information on the impact of dimensionality on the performance of the MH and LR procedures with real examinee data. A suggestion for further research is that the part dimensionality plays in DIF analysis may be further investigated with a simulation study so that the dimensionality of the test can be specified.

CHAPTER I

INTRODUCTION

Tests play an important role in today's society. It is vital, therefore, that test developers and users strive to ensure the validity of their instruments for the population and purposes for which they are intended. However, conscientious test developers and users have recognized that test scores were sometimes influenced by extraneous factors other than the construct a test was designed to measure. Because of this, differences in test scores do not always accurately reflect differences in examinees' true abilities. Whenever tests are designed for a population which displays gender, social, or culture differences, and whenever test results are used to inform decisions regarding placement, advancement, and competency, the possibility of the examinees' scores being influenced by biased test items is a matter of concern. Therefore, it is important to ensure that items which make up a test do not give undue advantage for success to members of one subgroup over another. Because of this, the detection of bias in testing has been a major issue in educational measurement. Bias can be assessed at the test level or at the item level. In this study, bias is addressed at the item level.

The study of test items that function differently for different groups of examinees was originally called "item bias" research (Holland & Thayer, 1986). Different groups of examinees may react differently to the same test questions (Holland & Thayer, 1986). It is necessary to investigate these differences since they may help to better understand both the characteristics of test items and the characteristics of different groups of

examinees in terms of their cognitive processes, test-taking strategies, and knowledge deficiencies.

The statistical process of investigating item bias involves gathering empirical evidence concerning the relative performance of different groups of examinees on a test item. However, statistical evidence of differential performance is necessary, but not sufficient, to draw the conclusion that bias is present in a test item. This conclusion involves an inference that goes beyond the data (Hambleton, Swaminathan, & Rogers, 1991). The empirical indices of bias do not necessarily indicate bias in the sense that it means unfairness. Groups may differ in their responses to an item for reasons other than bias, such as differential course taking patterns or interest in the specific content of the item (Donoghue & Allen, 1993). From empirical evidence, the only valid conclusion that could be made about an item is that it is functioning differentially in different groups of examinees. Therefore, more neutral and accurate terms such as differential item functioning (DIF) or differential item performance (DIP), rather than item bias, are preferred to describe the empirical evidence obtained in the investigation of bias (Hambleton, Swaminathan, & Rogers, 1991). In this study, the term DIF is used.

The psychometrically accepted definition of DIF is that an item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item correct (Hambleton, Swaminathan, & Rogers, 1991). This requires that individuals being compared are comparable with respect to the ability being measured by a set of test items or any other ability measure. If the performance on an item between two unmatched groups of examinees is being compared, the result may be

a measure of item impact rather than of DIF (Holland & Thayer, 1986). Impact is defined as the difference in performance that occurs only on the valid skill a test is designed to measure.

DIF can be further distinguished as uniform or nonuniform (Mellenberg, 1982). An item displays uniform DIF if there is no interaction between ability level and group membership and the probability of correctly answering the item is greater for one group than the other uniformly over the entire ability range (Swaminathan & Rogers, 1990a). That is, the advantage of one group over the other is consistent across ability levels. Nonuniform DIF exists when there is interaction between ability level and group membership; the difference in the probability of answering the item correctly for the two groups is not the same at all ability levels (Swaminathan & Rogers, 1990a); or two groups of interest are each favoured at certain ability levels. In terms of item response theory, uniform DIF is represented by parallel but not coincident item characteristic curves (ICCs) while nonparallel (ordinal interaction) or crossing ICCs (disordinal interaction) indicate nonuniform DIF.

Numerous statistical methods have been proposed for identifying test items that do not function in the same manner for specific subgroups. Among them, the Mantel-Haenszel (MH) procedure proposed by Holland and Thayer (1986) has emerged as one of the most popular procedures for detecting DIF because of its conceptual simplicity, ease of implementation, relatively small sample size requirement, and associated test of significance. However, the MH procedure is designed for detecting uniform DIF and may not be effective in detecting nonuniform DIF. As an extension of the MH procedure, a

more general DIF detection method, a logistic regression procedure, has been developed by Swaminathan and Rogers (1990a). This procedure has been shown to be powerful in detecting both uniform and nonuniform DIF (Ibrahim, 1992; Li & Stout, 1994; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990a, 1990b).

Detection of DIF for both the MH and LR procedures is based on chi-square tests of significance. But, each has its own advantage over the other. Swaminathan and Rogers (1990a) showed the connection between the two procedures. According to them, the MH procedure can be thought of as being based on a logistic regression model where the ability variable is considered discrete and no interaction between the ability variable and the group variable is permitted. Therefore the MH procedure is sensitive to only one type of DIF. While the LR procedure can be used to detect both uniform and nonuniform DIF, the MH procedure, on the other hand, provides a measure of the amount and direction of DIF.

Many researchers have studied the effectiveness and consistency of the MH procedure in the identification of DIF, but little has been done on the performance of the LR procedure. Moreover, in most of the studies on the performance of the MH and LR statistics, simulated data were used. Because of the high popularity of the MH procedure and potential usefulness of the LR procedure, and also because of the close connections between the two procedures, there is a need to further study and compare their performance under different conditions, especially with real examinee data from a published test.

In the following chapter, a review of the literature is presented.

CHAPTER II

REVIEW OF THE LITERATURE

In this chapter, a review of the literature is presented. This includes an overview of DIF detection procedures, a description of the Mantel-Haenszel and the logistic regression procedures, a review of selected studies examining the distribution and performance of the two statistics, and a summary of the findings. This leads to the purpose and specific research questions addressed in this study.

Overview of DIF Detection Procedures

The problem of assessing DIF has concerned test developers and measurement specialists for many years. Both judgemental and statistical approaches have been employed to examine it. So far, very little convergent validity is known to exist between results obtained from the two approaches (Sudweeks & Tolman, 1990). However, the two approaches should be used to complement each other.

In judgemental approaches, the test items are usually examined by expert judges in the appropriate knowledge area. The judges examine the item format and content for evidence that may cause the item to be more difficult for members of one group than for members of another group. The major problem with this procedure is the subjectivity of the expert judges.

Statistical procedures involve gathering empirical evidence of the performance of different groups of examinees on a test item. By using statistical techniques, researchers

attempt to sort out the difference between real performance and DIF by using the total test score or some other ability measure as the overall measure of ability or achievement level. Statistical procedures can be divided into conditional and unconditional procedures. In conditional procedures, the focus of analyses is on differences in performance between examinees from different groups of interest who are matched with respect to the ability, knowledge, or proficiency. In unconditional procedures, the examinee samples are not matched on any criterion. Unconditional procedures (such as Transformed Item Difficulty) have been criticized for confounding DIF with impact. The conditional procedures are more psychometrically acceptable because they incorporate the notion of comparing the item performance of members of different groups who are in some sense comparable. The conditioning or matching by ability is intended to produce an appropriate statistic that is sensitive to true differences in item functioning, provided the ability estimate accurately reflects the level of the true ability for these groups. If there is DIF under this control condition, then that difference in item performance is interpreted as a difference attributable to characteristics of the particular item and not due to difference in the characteristics of the individuals (Rock & Chan, 1988).

Many statistical methods have been proposed for detecting DIF, such as analysis of variance approaches, factor analytical techniques, distracter analysis, indices based on item difficulty, methods based on item response theory (IRT), and chi-square procedures. They differ from each other in terms of theoretical soundness, mathematical sophistication, sample size requirement, and ease of implementation. Over the years, attempts have been made to examine DIF detection procedures for their efficiency in

identifying items with DIF, and their reliability as determined by the degree to which a given procedure can replicate DIF identification results (Ibrahim, 1992).

Among the numerous DIF detection procedures, IRT-based methods have been theoretically preferred for the purpose of DIF identification. Most IRT-based methods involve a comparison of item characteristic curves for subgroups of interest. An item is judged to be unbiased if the item characteristic curves (ICCs) are the same for two groups of examinees. ICCs describe the relationship between the probability of a correct response and the examinee ability as measured by the test. In general, as ability increases, so does the probability of answering the item correctly. The ICCs may be defined by as many as three parameters, item difficulty (b), item discrimination (a), and probability of getting the item correct with zero knowledge (c).

While the IRT-based methods have theoretical appeal, difficult problems sometimes arise in applying them, especially when the three-parameter IRT model is used. Strict IRT assumptions, complex statistical calculations, high costs associated with running an IRT computer program such as LOGIST, large sample size requirement, and sometimes poor parameter estimates make the implementation of IRT-based methods problematic, if not impossible in some situations (Hambleton & Rogers, 1988). In addition, except for Lord's chi-square, most of the IRT indices do not have associated tests of significance. Therefore, a cutoff value must be determined in order to interpret DIF indices.

Because of the practical problems involved in using IRT methods, many researchers have sought alternative procedures in investigating DIF. Currently, two non-IRT procedures have been of great interest in research, the Mantel-Haenszel (MH) procedure

(Holland & Thayer, 1986) and the logistic regression (LR) procedure (Swaminathan & Rogers, 1990a).

The MH procedure is one of the most popular non-IRT DIF detection procedures. It has gained popularity on both theoretical and practical grounds. It helps to identify differences in performance on an item-by-item basis that may reflect irrelevant characteristics in certain test items that may be unfair to certain groups of examinees. In contrast to some of the IRT-based DIF detection procedures, the MH procedure is nonparametric, and it requires no model calibration (Ackerman & Evanness, 1992). The calculations are relatively simple. Computer programs for calculating the MH statistics are inexpensive to run. Also, the MH procedure requires sample size considerably below the size needed for the three-parameter IRT procedures in order to produce reliable results. Most important of all, there is an associated statistical test with a known sampling distribution (chi-square with one degree of freedom) and it tests the hypothesis of no bias (H_o) against a specific alternative hypothesis (H_1). In addition, the MH procedure provides a measure of effect size. Holland and Thayer (1986) examined the relationship between the MH and IRT-based procedures for assessing DIF. They concluded that the MH procedure compared favourably to the more complex and expensive IRT-based methods for detecting uniform DIF.

More recently, a potentially promising DIF detection technique, the logistic regression (LR) procedure, has been proposed by Swaminathan and Rogers (1990a). This procedure uses a standard logistic regression model for predicting a dichotomous dependent variable from given independent variables. As with the MH procedure, the LR procedure

uses a chi-square test of significance. According to Swaminathan and Rogers (1990a), the LR procedure takes into account the continuous nature of the ability scale and is useful for identifying both uniform and nonuniform DIF. Also, this procedure is more cost-effective to implement than IRT-based methods, although more expensive than the MH procedure.

The Mantel-Haenszel (MH) Procedures

According to Holland and Thayer (1986), the MH procedure was originally proposed by Mantel and Haenszel (1959) for the study of matched groups in the medical field. In this procedure examinees are first matched on a relevant criterion for the test under investigation. The most practical criterion is the total test score. However, an external criterion may be substituted. Based on total test score, score intervals or ability levels are formed for reference group and focal group examinees. The performance of the reference group is taken as a standard against which the performance of the focal group, which is of primary interest, is compared. All the items in a test are examined one by one. The item that is being analyzed for evidence of DIF is referred to as the studied item (Holland & Thayer, 1986). Then the data for the studied item for the examinees in the reference and focal groups is arranged into a series of contingency tables. There is one such table for each ability level (see Table 1), where one dimension of the table is the two groups being compared, and the other dimension is whether the examinees answer the studied item correctly or incorrectly.

Table 1: Data for the jth matched set of members of reference and focal groups (Holland & Thayer, 1986)

		Score on Studied Item 1	0	Total
Group	Reference	A_j	B_j	n_{Rj}
	Focal	C_j	D_j	n_{Fj}
	Total	m_{1j}	m_{0j}	T_j

Where: A_j and C_j denote the number of examinees in the reference and focal groups, respectively, in the jth score interval who answered the studied item correctly; B_j and D_j represent the number of examinees in the reference and focal groups, respectively, who responded incorrectly to the studied item; m_{1j} and m_{0j} are the total number of examinees who responded to the item correctly and incorrectly, respectively; n_{Rj} and n_{Fj} denote the number of examinees in the reference and focal groups respectively; T_j is the total number of examinees in the jth score group.

The MH-Alpha (α_{MH})

Based on the contingency tables, the ratio of the odds for success of the reference and focal group members (called alpha) is calculated for each score interval, weighted according to the number of individuals in that interval. These odds ratios are then averaged across the entire score scale to obtain the Mantel-Haenszel statistic.

The estimate of MH-Alpha, which expresses the common odds-ratio of success of the two groups across all score groups, can be defined as:

$$\alpha_{MH} = \frac{\Sigma A_j D_j / T_j}{\Sigma B_j C_j / T_j} \qquad (1)$$

This ratio is not symmetric. It has a scale of 0 to infinity, with $\alpha_{MH} = 1$ representing the null value of no DIF. For a given item, it is the weighted average of the odds ratios taken across all score intervals. It is the average amount by which the odds that a reference group member is correct on the studied item are greater than the corresponding odds for a member of the focal group with the same ability. When the odds for success on an item of the reference and focal groups among examinees of the same ability levels are substantially different, DIF is suspected. When $\alpha_{MH}=1$, the odds for success are the same in the reference and focal groups. Then no DIF has been detected. The item is equally difficult for both groups. When $\alpha_{MH}>1$, the item appears to be functioning differentially in favour of the reference group. When $\alpha_{MH}<1$, the focal group performed better on average than did the reference group (Holland & Thayer, 1986).

The MH Chi-Square (MH-CHISQ)

Associated with α_{MH} is a chi-square test of significance. Let p_{Rj} and q_{Rj} represent the probabilities of answering the item correctly and incorrectly, respectively, for members of the reference group in the jth interval, and p_{Fj} and q_{Fj} denote the corresponding probabilities for the focal group members, the MH chi-square tests the null hypothesis:

$$H_o : \frac{p_{Rj}/q_{Rj}}{p_{Fj}/q_{Fj}} = 1 \qquad \text{for all } j=1,\ldots,k \qquad (2)$$

versus

$$H_1 : \frac{p_{Rj}/q_{Rj}}{p_{Fj}/q_{Fj}} = \alpha \qquad \text{for } \alpha \neq 1,\ j=1,\ldots,k \tag{3}$$

The hypothesis being tested here is that the sum of all the odds ratios across the matched sets (α_{MH}) for a given item are unity. The MH chi-square for testing the hypothesis that $\alpha_{MH}=1$ has the form:

$$\text{MH-CHISQ} = \frac{(\,|\,\Sigma_j A_j - \Sigma_j E(A_j)\,|\,-1/2)^2}{\Sigma \text{Var}(A_j)} \tag{4}$$

where

$$E(A_j) = n_{Rj} m_{1j} / T_j \tag{5}$$

and

$$\text{Var}(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)} \tag{6}$$

While α_{MH} indicates practical difference between the reference and focal groups, the MH chi-square statistic indicates whether there exists a statistically significant difference between the comparison groups. The MH chi-square has an approximate chi-square distribution with one degree of freedom under the null hypothesis of no DIF on the studied item for matched reference and focal group members. The alternative hypothesis is nondirectional. Consequently, the MH chi-square statistic can identify DIF that favours either subgroup (Holland & Thayer, 1986). According to Holland and Thayer (1986), a test based on MH chi-square is the uniformly most powerful unbiased test of H_o versus H_1.

The MH-Delta (Δ_{MH})

Holland and Thayer (1986) proposed a log transformation to put α_{MH} on the ETS (Educational Testing Service) delta scale. The ETS delta scale is a normalizing transformation of the item difficulty. The item difficulty values (as measured by the proportion of correct responses) are first normalized for each group by converting to z scores corresponding to the (1-p)th percentile. The z scores are adjusted by a linear transformation to delta values with a mean of 13 and standard deviation of 4 (Wright, 1986). A log transformation can be used to put α_{MH} on the ETS delta scale with zero as the null value:

$$\Delta_{MH} = -(4/1.7)\ln(\alpha_{MH}) = -2.35\ln(\alpha_{MH}) \quad (7)$$

This transformation results in a negative value for items that are more difficult for the focal group. The value of Δ_{MH} is the average amount more difficult that a member of reference group found the studied item than did comparable members of focal group. For items of average difficulty, a difference of 1 delta unit is equivalent to approximately a .10 difference in item difficulty and a difference of .8 delta unit is equivalent to approximately a .08 difference in item difficulty. A negative value of Δ_{MH} indicates that an item favours the reference group, a positive value indicates that an item is easier for the focal group. Δ_{MH} is a measure of the size and direction of DIF in the scale of differences in item difficulty between the reference and focal groups as measured in the ETS delta scale (Holland & Thayer, 1986).

Δ_{MH} is very similar to the effect size measure in analysis of variance. According to

ETS criteria, test items can be classified into three categories based on the absolute values of Δ_{MH} and their significance level (.05) (Zwick & Ercikan, 1989):

"A" items are those that have absolute Δ_{MH} values less than 1 or not significantly different from zero ($p<.05$). These items are considered to function properly or free of DIF.

"B" items are those that have Δ_{MH} values that are significantly different from 0 ($p<.05$) and have either absolute values of Δ_{MH} at least one but less than 1.5 or absolute values of one but not significantly greater than one ($p<.05$). These items may also be used; however, among otherwise equivalent items, those with the smallest values of Δ_{MH} are preferred.

"C" items are those with absolute Δ_{MH} values at least 1.5 and are significantly greater than one ($p<.05$). These items are considered to be showing DIF. They should be included in a test only if it is essential to meet test specifications.

Program MANTEL developed by Ackerman (1987) for caiculating the MH statistics produces values of MH-Z, which is another linear transformation of α_{MH}.

$$MH\text{-}Z = \ln(\alpha_{MH})/(-1.7) \qquad (8)$$

MH-Z and Δ_{MH} have basically the same implication, only expressed on different scales. Both have a value of 0 under the null hypothesis and both provide a measure of the amount and direction of DIF. Actually, Δ_{MH} is equal to MH-Z multiplied by 4. For items of average difficulty, a difference of .25 MH-Z is equivalent to approximately a .10 difference in item difficulty and a difference of .20 MH-Z is equivalent to approximately a .08

difference in item difficulty. Since this is the only available computer program for calculating the MH statistics, MH-Z instead of Δ_{MH} is used in this study. In a sense, the MH-Z is better than Δ_{MH} because item difficulty is expressed in standard z scores.

Two steps are suggested by Holland and Thayer (1986) for calculating the MH statistics:

(1) The MH statistics are computed for all test items with the score groups constructed using total scores based on all items. The resulting statistics are used as an initial criterion to identify DIF.

(2) All items showing DIF are eliminated from the test except one (the studied item) and the total test score of the remaining items plus the studied item is computed again, and used as the new matching criterion. Score groups are reformed, and the MH statistics are recalculated for all items.

The Logistic Regression (LR) Procedure

The LR procedure provides a model-based approach for studying DIF. The standard logistic regression model for predicting the probability of a correct response to an item from given independent variables is

$$P(u = 1 \mid \theta) = \frac{e^{(\beta_0+\beta_1\theta)}}{[1 + e^{(\beta_0+\beta_1\theta)}]} \qquad (9)$$

where u is the response to the item, θ is the observed ability of an individual, β_0 is the intercept parameter, and β_1 is the slope

parameter.

The logistic regression model given above can be used to model differential item functioning by specifying separate equations for the focal and reference groups:

$$P(u_{ij} = 1 \mid \theta_{ij}) = \frac{e^{(\beta_0+\beta_1\theta_{ij})}}{[1 + e^{(\beta_0+\beta_1\theta_{ij})}]}, \quad i=1,...,n_j,\ j=1,2. \tag{10}$$

where u_{ij} is the response of individual i in group j to the item, θ_{ij} is the ability of individual i in group j, β_{0j} is the intercept parameter and β_{1j} is the slope parameter for group j.

This model is used to predict the probability of a correct response to an item as a function of examinee ability and group membership. When $\beta_{01}=\beta_{02}$ and $\beta_{11}=\beta_{12}$, the logistic regression curves for the two groups are the same, and there is an absence of DIF. An item shows uniform DIF if $\beta_{01}\neq\beta_{02}$ while $\beta_{11}=\beta_{12}$, in which case, the logistic regression curves are parallel but not coincident. The presence of nonuniform DIF may be inferred if $\beta_{11}\neq\beta_{12}$ whether $\beta_{01}=\beta_{02}$ or $\beta_{01}\neq\beta_{02}$ (Swaminathan & Rogers, 1990b).

Swaminathan and Rogers (1990a) stated that the estimates of the parameters (β_{01}, β_{02}, β_{11}, and β_{12}) are asymptotically normally distributed. As the MH procedure, the LR procedure also uses a chi-square test of significance to test the hypothesis of no DIF, which is

$$H_o: \beta_{01}=\beta_{02} \text{ and } \beta_{11}=\beta_{12}$$

The resulting statistic is a chi-square with two degrees of freedom. In this case, the LR procedure does not differentiate between uniform and nonuniform DIF.

The LR procedure can also be used to test the hypothesis of no DIF against the hypothesis of uniform DIF and the hypothesis of nonuniform DIF separately. The results are two chi-square statistics (one for uniform DIF and one for nonuniform DIF), each with one degree of freedom.

Selected Studies Examining the Distribution and Performance of the MH and LR Procedures

Since Holland and Thayer (1986) introduced the MH approach, many researchers have studied the effectiveness of this method in identifying differentially functioning test items under different conditions. Some researchers compared its performance with other accepted DIF detection procedures. Others studied its stability by manipulating variables of interest such as sample size, item discrimination, item difficulty, number of score groups, ratio of reference to focal group members, nature of DIF (uniform vs. nonuniform), and the criterion for measuring ability.

The influence of sample size on the functioning of the MH statistics was studied by Mazor, Clauser, and Hambleton (1991). Three data sets of item responses of 2,000 examinees each were generated using a three-parameter IRT model with "c" parameter fixed at .20. For two of these sets, the ability scores were set to have a normal distribution with a mean of 0 and standard deviation of 1. These were referred to as Reference Group 1 and Focal Group 1 with equal abilities. The ability for the third distribution was set to a mean of -1 and standard deviation of 1. This was referred to as

Focal Group 2 which was compared with Reference Group 1 (unequal ability groups comparison). Five sets of 16 DIF items each (80 DIF items) were generated and combined separately with 59 non-DIF items to create five different 75-item tests. The 80 DIF items reflected five levels of "b" (-2.5, -1.0, 0, 1.0, and 2.5), four levels of "a" (.25, .60, .90, and 1.25) and four levels of difference in "b" (.25, .50, 1.0, and 1.5) between reference and focal groups.

The MH procedure was run for each test comparison. Tests were first compared using all 2,000 examinees in each group. The first 1,000 examinees in each group were then selected, and the procedure was rerun. Similarly, this was repeated with 500, 200, and 100 examinees. To minimize the impact of chance variability, for sample size at 500 one replication was done for each set, and for sample size at 200 and 100 two replications were run.

The results showed that the percentage of DIF items correctly identified decreased as the number of examinees decreased. The MH procedure identified 74% and 64% of the DIF items (for the equal and unequal distributions, respectively) when groups of 2,000 were used. With a sample size of 1,000, 61% and 58% (for the equal and unequal distributions, respectively) of the DIF items were detected. With 500 or fewer examinees, more than 50% of the DIF items were not identified. The high percentage of items missed even at the largest sample sizes (2,000) was unexpected. The items most likely to be missed by the MH procedure in this dataset were those that were very difficult (b=2.5), those with a small difference in "b" parameter (.25) between the two groups, and poorly discriminating items (a=.25). When the ability distributions for the groups were

equal, the detection rates were consistently higher than when the distributions were unequal.

The differences between the p-values (classical definition of item difficulty) of reference and focal groups were also calculated based on the sample size of 2,000. A comparison of the p-value differences between groups, associated with items identified or missed by the MH procedure, suggested that the statistic might have greater sensitivity when used with groups of equal ability than when used with groups of unequal ability. The largest p-value differences associated with DIF items that were missed with the equal ability comparison were .04, .08, .08, .17, and .23 when using groups of 2,000, 1,000, 500, 200, and 100, respectively. This was in contrast to .07, .15, .17, .23, and .29 when using groups of 2,000, 1,000, 500, 200, and 100, respectively, for the unequal ability comparison.

Some of the results of the above study are very similar to those of a study by Clauser, Mazor, and Hambleton (1991a) in which they used simulated data with the same item parameters to examine the effects of item discrimination and difficulty parameters on the MH procedure. They found that the MH chi-square values increased as the difference in the difficulty parameter values between the reference and focal groups increased, but the size of this increase was clearly related to item discrimination. Low discriminating items (a=.25) were associated with low MH chi-square values even when the difference in the difficulty parameters was large. Highly discriminating items ($a \geq .90$) produced high MH chi-square values when compared across similarly divergent difficulty parameter values. In addition, DIF in very difficult items (b=2.50) was not identified by the

MH procedure indicating the procedure did not function equally effectively across the entire range of the difficulty scale. This is because there were too few examinees functioning in that part of the ability distribution. Since the MH statistics are weighted by the number of examinees in each ability level, DIF occurring within a few sparsely populated score intervals at the extreme end of the distribution might be missed.

In another study by Clauser, Mazor, and Hambleton (1991b), the influence of the criterion variable on the identification of differentially functioning test items using the MH statistic was examined. The dataset for the study was from the 1982 administration of the New Mexico High School Proficiency Exam which included the responses of 23,000 students to 150 test items. Two random samples of 1,000 Anglo-American and 1,000 Native American examinee item response sets were selected and analyzed. Only 91 of 150 items were used after removing items which discriminated poorly as measured by the biserial correlation ($r<.1$), or were very easy ($p>.90$) in the combined sample of 2,000 examinees. The 91 items were assigned to three groups of 75 each with the constraint that each item be included in at least two groups. Therefore, there were 43 items in common in the three tests. The Mantel-Haenszel procedure was first applied to the three 75-item tests. Items were identified as DIF only if they were identified as DIF in each of the tests in which they appeared. The items were then categorized as belonging to one or more of four subtests based on the skills or knowledge needed to select the correct response (math, reading, prior knowledge, and charts). The first three categories are mutually exclusive. Items in the "charts" category also appeared in one of the other three categories. The Mantel-Haenszel statistics were then recalculated for items in each of the

four subtests. Three additional subtests were also constructed by randomly assigning each of the 91 items to one of the three groups (30, 31, 30) and were analyzed using the Mantel-Haenszel computer program. The same set of examinees (1,000 Anglo-American and 1,000 Native American) were used in all of the above analyses.

The results showed that based on three runs of the MH computer program analyzing a total of 91 items (randomly assigned to three 75-item tests with the constraint that each item be included in at least two groups), 22 items were identified as showing differential item functioning. When the original set of items analyzed was regrouped and reassessed within four separate subtests, 32% of the DIF items (7 of 22) were no longer found to be differentially functioning and 11 new items were identified as DIF. A comparison of DIF and Non-DIF items in the total test and subtests suggests that systematically changing the grouping of the test items analyzed resulted in some items showing substantial changes in their MH statistics. The results indicated that the choice of criterion, total test score versus subtest score, had a substantial influence on the classification of items as to whether or not they were differentially functioning between the Anglo- and Native American groups.

Using the same dataset, Hambleton and Rogers (1988) conducted a detailed analysis to examine the degree of agreement between the IRT Area method and the Mantel-Haenszel method in identifying DIF items. Another purpose of their study was to examine the performance of the two procedures when the ability distribution of the two groups of interest was considerably different. Four samples of 1,000 examinees each were drawn (2 Anglo-American and 2 Native American groups). Only 75 of the original 150 items were

used. As in the study by Clauser et al. (1991b), items with low discrimination (biserial correlation r<.10) or which were very easy (p>.90) were excluded because these items might cause difficulties in the IRT parameter estimation and lead to unstable DIF statistics. Three-parameter IRT models were fitted to each of the four samples separately. With 2 Anglo-Native American samples, one replication was conducted to enable the examination of the consistency with which each DIF statistic identified items across samples. The MH and ICC (item characteristic curve) Area statistics were calculated for each item in each of the two comparisons. For the ICC Area method, a cutoff value of .468 was obtained by carrying out an analysis on two randomly equivalent groups (the two Native American samples).

The IRT Area method results were consistent across samples (of 1,000) about 73% of the time and the MH results were consistent about 80% of the time. Of the 14 items consistently identified by the Area Method across the two comparisons and the 9 items consistently identified by the MH method, 7 items were common. Those items identified by the MH procedure were more or less a subset of those identified by the Area Method. The items not identified by the MH procedure were mostly items with nonuniform DIF. The authors concluded that there is substantial agreement between the IRT Area method and the Mantel-Haenszel procedure in the detection of items with uniform DIF. Also, the IRT Area method appeared to detect items with nonuniform DIF; the Mantel-Haenszel did not.

To study the effect of the score distribution on the MH method statistics, a matched group analysis (650 Native and 650 Anglo-Americans) was carried out. Both the MH and IRT Area statistics were calculated. The ability interval over which the area was calculated

was modified to cover the ability scale from 2 standard deviations below the Native American group mean to 2 standard deviations above the same mean. Thus, attention was focused on the part of the ability scale where most of the Native American examinees were located. Restricting the ability interval to focus on the region where most of the focal group was distributed led to the identification of fewer non-uniformly biased items with the Area method, and hence, greater congruence with the Mantel-Haenszel results. The distribution of test scores appeared to have little impact on the Mantel-Haenszel statistics. Matching the groups according to test score distribution before calculating the DIF indices did not substantially change the results. The Area method results were influenced to a much greater extent, although this may have been due in part to the reduction in sample size which was necessary to achieve matching.

Hambleton, Rogers, and Arrasmith (1988) compared the Mantel-Haenszel statistic with three other DIF detection methods: the weighted b-value plot method, the root mean squared difference method, and the total area method. The latter three methods are based on item response theory. The test data consisted of the item responses of 937 ninth grade students to the 75 test items on the 1985 Cleveland Reading Competency Test. Simulated data were used to set cutoff scores for interpreting IRT DIF statistics. For the MH statistic, a .01 level of significance was used. For the weighted b-value plot method, root mean squared difference method, and the total area method, cutoff values of 1.40, .11, and .50 respectively were used. The number of items identified varied from 5 to 8 across the methods. There was moderate agreement for the items identified as showing DIF by the four methods. The evidence suggests that the four methods, leaving

methodological problems aside, led to the identification of nearly the same set of items as showing DIF. The four methods did not differ in their detection of DIF items by more than one item or two after methodological shortcomings were taken into account. The choice of interval over which bias was defined appeared to be the cause of differences. Methodological problems included imprecision in establishing cutoff points, Type I errors, and poor item parameter estimates. The authors concluded that the MH statistic provided a quick, cheap alternative to the more laborious and expensive item response theory methods.

Many additional studies have been done on the MH procedure demonstrating it's validity. However, very few studies relating to the LR procedure have appeared in the literature. The use of a LR procedure has only recently been proposed by Swaminathan and Rogers (1990a). So it is relatively untried. However, since it can be used to identify both uniform and nonuniform DIF, it has already attracted researchers' attention.

Using simulated data, Swaminathan and Rogers (1990b) compared the power and detection rates of the MH and LR procedures in identifying items showing DIF. The factors manipulated in the study were sample size (250 per group/500 per group), test length (40 items/60 items/80 items), and the nature of the DIF (uniform/nonuniform). The two DIF detection procedures were compared with respect to the percentage of items with uniform and nonuniform DIF that was correctly identified and the percentage of items with no DIF that was falsely identified. In addition, the power of each DIF detection procedure was studied by carrying out twenty replications on the combination of sample size of 500 per group and test length of 80 items. No replications were done for other sample sizes

or test length.

The results showed that for the items with uniform DIF, the two procedures had very similar detection rates, with the MH procedure having a slight advantage. Both procedures were able to detect uniform DIF with about 75% accuracy in samples of 250 per group and with 100% accuracy in samples of 500 across the three levels of test length. For nonuniform DIF, the MH procedure completely failed to detect nonuniform DIF under any condition. The LR procedure detected nonuniform DIF with about 50% accuracy with small samples (250) and short test length (40 items) and 75% accuracy with large samples (500) and long test length (80 items). The MH procedure consistently produced around 1% false positives under all conditions; the LR procedure produced between 1% and 6% false positives. Results of the power study showed that the two procedures are similarly and highly powerful in detecting uniform DIF, but only the LR procedure is able to detect nonuniform DIF.

Rogers and Swaminathan (1993) examined the distributional properties and the power of the LR and MH statistics. In the distribution study, four conditions were simulated by crossing two levels of model-data fit (good and poor fit) with two levels of sample size (250 per group/500 per group). Data for a 40-item test for which the LR model provided good or poor fit were generated using the two-parameter logistic IRT model and the three-parameter logistic model, respectively. One hundred replications were conducted for each combination of sample size and model-data fit. Five of the 40 items were chosen to vary in level of difficulty ("b"=-1.5, "b"=0, and "b"=1.5) and discrimination ("a"=.6, "a"=1, and "a"=1.6). For each of these five items, the LR and MH statistics were calculated, and

empirical sampling distributions were constructed and tested by the Kolmogorov-Smirnov test to determine if the statistics had the expected distributions.

The results of the distribution study showed that for both the LR and MH test statistics, the expected distributions were obtained under nearly all conditions. The LR test statistic did not have the expected distribution for very difficult and highly discriminating items.

In the power study, 32 conditions were simulated by crossing two levels of model-data fit (good and poor fit), two levels of sample size (250 per group/500 per group), two levels of test length (40 items and 80 items), two levels of the shape of the test score distribution (normal and negatively skewed), and two levels of percent of items with DIF. Each condition was replicated 20 times. With each condition, both uniform and nonuniform DIF were simulated by varying "a' and "b" parameters. In simulating uniform DIF, the "a" parameters for the two groups were kept the same but the "b" parameters for the two groups were different. In simulating nonuniform DIF, the "b" parameters for the two groups were the same, but the "a" parameters for the two groups were different. For each type of DIF in each condition, four sizes of DIF were studied, corresponding to IRT area values of .2, .4, .6, or .8. Finally, the data were analyzed by ANOVA in which the dependent variable was the number of times out of the 20 replications that the item with DIF was identified by each procedure, and the independent variables were the factors manipulated.

For uniform DIF, the results of the power study showed that test length and shape of the score distribution did not affect detection rates of either procedure. Model-data fit did

not affect the LR procedure but had a significant effect for the MH procedure. The MH procedure was not influenced by the percent of DIF items but the LR procedure was. Sample size had a strong effect on the detection rates for both procedures. Detection rates increased by 15% when sample size was increased from 250 to 500. The type of item also had a large effect for both procedures. The items that were most easily detected were items of moderate difficulty and high discrimination. Size of DIF produced the expected effect, detection rates for both procedures were low for DIF items with an area value of .2, but were high for DIF items with an area value of .6. Overall, the LR and MH procedures were almost equally effective in detecting uniform DIF with the MH procedure having a slight advantage.

Results for the detection of nonuniform DIF showed that both procedures were unaffected by the shape of the score distribution or the percent of items with DIF. The MH procedure was not sensitive to test length, but the LR procedure was slightly affected. Both procedures were affected by model-data fit. The detection rates for both procedures were lower when the model-data fit was poor than when it was good. Again, sample size produced a strong effect. Other than the size of DIF, the largest effect observed for both procedures was due to the type of item. For the LR procedure, the lowest detection rate occurred with items of moderate difficulty and low discrimination, and the highest detection rate occurred for items of moderate difficulty and high discrimination. The MH procedure was unable to detect nonuniform DIF in items of moderate difficulty. For items of low difficulty, the MH detection rate was still 15% lower than the LR detection rate; for items of high difficulty, the detection rates were nearly the same. The authors concluded

that the LR procedure was as powerful as the MH procedure in detecting uniform DIF, and more powerful than the MH procedure in detecting nonuniform DIF.

Ibrahim (1992) examined the distributional properties and the power of several DIF identification procedures including the MH chi-square, the MH-Z, the LR, etc. Item response strings were simulated based on the three-parameter IRT model with the guessing parameter fixed at .15. Examinee samples were generated with abilities from a normal distribution (0,1). Two test lengths (42 items and 66 items) and four arrangements of "a" values were used. In the first arrangement (denoted as D1), all the items in the test were given an "a" value of .80. In the second arrangement (D2), the first half of the test had "a"=.80 while the second half of the test had "a"=1.20. For the third arrangement (D3), the first half had "a"=1.20 while the second half had a=.80. In the fourth arrangement (D4), all the items had "a"=1.20. "b" values set to range from -2 to 2 were classified into low "b", medium "b", and high "b". Sample sizes of 1,600/600, 800/300, 400/150 examinees with a ratio of 8:3 in the majority and minority groups, respectively were used. Within each experimental condition, 100 replications were performed. For each DIF index within each experimental condition, percentiles of .99, .95, and .90 for each item over 100 replications were computed. Multivariate analysis of variance and post hoc univariate ANOVA were used to test for differences in the mean values of the percentiles obtained over the 100 replications in each test across each experimental condition.

Under the null hypothesis, the distributions of the MH chi-square and the LR nonuniform chi-square DIF indices were observed to have sample size and test length

effects. Generally speaking, increasing sample size and test length resulted in an increase in the means of the three percentile values (.99, .95, and .90) and they were greater than expected. The distribution of the LR uniform chi-square indices was not affected by sample size and test length. The MH-Z was affected by item discrimination. High discrimination values were associated with high MH-Z values when only D1 and D4 were considered.

In the power study, six DIF items (3 with uniform DIF and 3 with nonuniform DIF) were simulated by altering either b-values or their a-values for the minority group. The LR chi-square and the MH chi-square statistics were equally powerful in detecting items with uniform DIF. Each of them had a detection rate of .96. For nonuniform DIF, the LR procedure was more powerful. It produced a detection rate of .87. The MH chi-square produced a rate of .56. The overall false-positive rate for the MH chi-square was .095. The false-positive rates for the LR chi-square were .087 and .076 for uniform and nonuniform DIF respectively.

With uniform DIF items, the detection rates of the MH chi-square decreased as levels of "b" decreased. When the items had nonuniform DIF, a much lower detection rate was obtained for items with medium "b". For items of medium difficulty, the ICCs for the two groups crossed in the middle of the ability range and the positive and negative values of the area between the ICCs would cancel each other; the MH chi-square was not sensitive to this kind of nonuniform DIF. Similarly, Swaminathan and Rogers (1990a) found that the MH procedure was insensitive to this kind of DIF. However, the MH chi-square picked up items with nonuniform DIF with low and high "b"s, in which case the DIF may appear to

be uniform over most of the ability range. The false-positive rates of the MH chi-square for low, medium, and high "b"s were .075, .111, and .094 respectively. For both uniform and nonuniform DIF, the detection rates of the LR procedure increased as levels of "b" decreased.

For uniform DIF, the MH chi-square had detection rates above .91 at the three sample sizes (1,600/600, 800/300, and 400/150); while for nonuniform DIF, the rates were .68, .57, and .42 for the sample sizes of 1,600/600, 800/300, and 400/150 respectively. False-positive rate of the MH chi-square also increased as sample size increased. The LR procedure also produced positive relationships between sample size and DIF detection rate for both uniform and nonuniform DIF. The false-positive rates for sample sizes of 800/300 and 400/150 were equal at .078. The false-positive rate was higher for sample size of 1,600/600 than those observed for the other two sample sizes by about .02.

Item discrimination and test length displayed no significant effect on the DIF detection rate of the MH, LR uniform and LR nonuniform chi-square statistics. But the false-positive rate was smaller for the longer test than for the shorter test.

Pang and Boss (1993) studied the effects of sample size, item discrimination "a", and item difficulty "b" on the distribution of the logistic regression indices under the null hypothesis. A 41-item test was simulated. Nine conditions were created by crossing three levels of sample size (1,000, 500, and 250 in both reference and focal groups) with three levels of "a" (.7, 1.0, and 1.3). Three levels of "b" values were used: low "b" (-2.0 to -.7), medium "b" (-.6 to .6), and high "b" (.7 to 2.0). The logistic regression uniform and nonuniform DIF indices were calculated and one hundred replications were done for each

condition. For each replication, the mean, standard deviation, and skewness of the distribution of the LR indices were computed for each experimental condition. In addition, three chosen cutoffs (P_{90}, P_{95}, P_{99}) were computed for each condition. A MANOVA was conducted on the mean, standard deviation, and skewness of the LR indices.

The results indicated that for uniform DIF, the distribution of the LR chi-square index was not affected by sample size, level of "a", and level of "b". However, the mean of the LR uniform chi-square DIF index was slightly underestimated compared to the expected values of a chi-square distribution. Contrary to Ibrahim (1992), sample size had no effect on the distribution of the LR nonuniform DIF indices. However, level of "a" and level of "b" affected the distribution of the LR nonuniform DIF index. Large values of "a" and "b" increased the mean and standard deviation. For high "b"s and "a"=1.3 the mean values exceeded the expected value of 1.0. For the standard deviation, underestimation was observed at low "b" and medium "b". Similar patterns were observed for skewness with both uniform and nonuniform DIF.

With uniform DIF, sample size, level of "a", and level of "b" did not significantly affect the three cutoffs (P_{90}, P_{95}, and P_{99}). However, the critical values were slightly overestimated at P_{90} and P_{95} and underestimated at P_{99}. With higher values than expected, a danger may exist that more biased items would be detected when, in fact, they are not biased. The authors concluded that the LR chi-square indices for uniform DIF did not fully fit the expected chi-square distribution in this dataset.

With the nonuniform DIF index, high discrimination value reflected higher values for P_{90} and P_{95} than expected. A significant effect was found for "a"=1.3 at P_{90} and P_{95}. As

was the case with uniform DIF, the critical values were overestimated for P_{90} and P_{95}. Thus, for both uniform and nonuniform DIF, Type I error may be greater than .05 and .10 if one chooses using P_{90} and P_{95} of the tabled values of the chi-square statistics as cutoffs.

Ochieng (1992) conducted a similar simulation study, and studied the effects of sample size, level of "b", level of "a", and ability distribution on the distributions and percentiles (P_{90} and P_{95}) of LR and MH indices under the null hypothesis. Item response strings were simulated for focal and reference groups for both equal and unequal ability distributions.

Consistent with the findings by Pang and Boss (1993), LR uniform DIF index was not influenced by sample size, level of "a", and level of "b". Consistent with Ibrahim (1992), sample size showed a significant effect on the MH chi-square and the LR nonuniform DIF indices. Ability distribution showed no significant effect on any of the four indices. Levels of "a" significantly affected the mean of the MH chi-square resulting in greater values at a high level of "a" than at a low level of "a". In the case of the LR uniform indices, "a" and "b" values had no significant effect, while for the LR nonuniform indices, levels of "b" showed a significant effect on the mean and standard deviation. For the LR uniform indices the means were close to the expected value of one while the standard deviation was slightly lower than the expected values. For the LR nonuniform indices both the means and standard deviation were both close to the expected values.

For LR uniform index, "a" values affected P_{90} and P_{95}. Level of "b" showed a significant effect on P_{90} but not on P_{95} of MH chi-square. For LR nonuniform index, levels

of "b" showed a significant effect on both P_{90} and P_{95}.

At a low sample size and high discrimination level the P_{90} of MH chi-square was smaller than the expected value of 2.70. However, at any other level of the independent variables, the P_{90} values of MH chi-square were overestimated suggesting that the use of tabled values would result in higher false positives than expected. The P_{90} and P_{95} obtained for LR uniform were larger than the expected values of 2.70 and 3.84 respectively. The P_{90} and P_{95} values were greater than the tabled values at all the levels of independent variables for LR nonuniform suggesting that more false positives would result when the tabled values are used.

Using real data from a Grade 4 provincial reading assessment test in British Columbia, Brown (1992) compared the performance of the MH and LR procedures in detecting DIF across sample sizes and over replications. The test consisted of two short subtests. Sixteen comparison groups were created by randomly selecting cases by gender and/or region across six levels of sample size: 1,000/1,000, 750/750, 500/500, 300/300, 200/200, and 100/100 examinees per group. The MH chi-square and LR uniform and nonuniform chi-square indices were then computed separately for each of the two subtests. Replications were carried out at each sample size from 1 to 5 times, depending on population available.

The results indicated that compared to the prespecified standards, the DIF detection rates were low for the MH chi-square, LR uniform, and nonuniform chi-square statistics, with the LR uniform procedure performing slightly better than the MH procedure. The false positive rate was about 6% for all three procedures. Agreement between the MH chi-

square and the LR uniform chi-square procedures were high. Sample size was found to significantly affect the detection rates of MH and LR. Detection rates were higher for large sample sizes.

Summary of the Findings

The Distribution Studies

Overall, the distributions of the MH and LR procedures appeared to be met to a satisfactory degree, at least for practical purposes (Rogers & Swaminathan, 1993). However, the distributions of the MH and LR nonuniform chi-square DIF indices were observed to have a slight sample size effect. Ibrahim (1992) indicated that increasing sample size resulted in greater MH and LR indices than expected. In Rogers & Swaminathan (1993), the distributional assumptions were less often met when the sample size of 250 was used than when the sample size of 500 was used. However, sample size did not affect the distribution of the LR uniform index (Ibrahim, 1992; Ochieng, 1992; Pang & Boss, 1993).

Test length influenced the distributions of the MH and LR nonuniform indices in that increasing test length resulted in a slight increase in the means of the three percentile values (.99, .95, and .90) (Ibrahim, 1992). The LR test statistic did not have the expected distribution for very difficult and highly discriminating items (Rogers & Swaminathan, 1993). In Ochieng (1992) and Pang and Boss (1993), item difficulty and discrimination did not affect the distribution of the LR uniform index under the null hypothesis, but the mean and standard deviation of the LR nonuniform index were slightly overestimated as the

values of item difficulty and discrimination became larger. Ochieng (1992) indicated that ability distribution had no effect on the distribution of the MH and LR indices.

The Power Studies

Sample size influenced the power of the MH statistics (Mazor, et al., 1991; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990a, 1990b) and the LR procedure (Ibrahim, 1992; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990a, 1990b). The power of the MH and LR statistics improved with larger samples.

Test length displayed no significant effect on the detection rates of the MH and LR procedures (Ibrahim, 1992; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990a, 1990b). However, Rogers and Swaminathan (1990a, 1990b, 1993) also indicated that for nonuniform DIF the detection rates of the LR procedure were slightly higher for longer tests than for shorter tests.

Item difficulty and item discrimination influenced the MH or the LR statistics (Mazor, et al., 1991; Swaminathan & Rogers, 1990a). In Ibrahim (1992), item difficulty affected the detection rates of the MH and LR chi-square statistics, but item discrimination did not. Poorly discriminating and very difficult items were less likely to be identified by both the MH and LR statistics except at large sample sizes.

The power of the MH and LR statistics to detect the presence of DIF was affected by the nature of DIF (uniform or nonuniform). Comparative studies between the MH and LR procedures showed that the LR procedure is as powerful as the MH procedure in identifying uniform DIF and more powerful than the MH procedure in identifying

nonuniform DIF (e.g. Brown, 1992; Ibrahim, 1992; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990a, 1990b).

The probability of success on an item depends on the examinee's ability. Disparity in ability distribution between the reference and focal groups was noted to influence the power of the MH procedure although little is known on the possible effects on the LR procedure (Clauser, et al., 1991; 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. However, Rogers and Swaminathan (1993) showed that the detection rates of the MH and LR procedures were not affected by the shape of score distribution. Hambleton and Rogers (1988) also indicated that the ability distribution did not affect the performance of the MH indices in their study.

The choice of criterion as a measure of ability, total test score or subtest score, had a substantial influence on the classification of items as showing DIF (Clauser, et al., 1991).

The MH statistics have been compared with the LR procedure (Rogers & Swaminathan; Swaminathan & Rogers, 1990a, 1990b) and some other DIF detection procedures such as IRT-based methods (e.g. Hambleton & Rogers, 1988; Hambleton, et al., 1988) and were found to be in agreement with the LR and other DIF detection methods except in the identification of nonuniform DIF. In identifying uniform DIF, the MH was found to be very powerful. The LR was found to be effective in detecting nonuniform DIF with relatively high accuracy (Rogers & Swaminathan, 1993; Swaminathan & Rogers,

1990a, 1990b).

Purpose of the Research

Both the MH and LR procedures have desirable statistical properties in identifying DIF. Many researchers have studied the effectiveness and consistency of the MH procedure in the identification of DIF, but little has been done on the performance of the LR procedure. Moreover, in most of the studies described above on the distribution and power of the MH and LR statistics simulated data were used. One advantage of a simulation study is that items with or without DIF and the nature of DIF are generated; thus, the true DIF conditions of items are known. The power and detection rates of different procedures can then be compared to the known DIF conditions. Also, the Type I error rates of the DIF detection statistics can be assessed. Another advantage is that the performance of the DIF detection procedures can be investigated under different conditions by manipulating independent variables such as sample size, item difficulty, item discrimination, and ability distribution. The third advantage is that substantially large numbers of replications can be carried out, which makes it easier to assess the stability of DIF indices.

However, the limitation of a simulation study is that the data are generated, and they may not reflect real life situations. Results based on simulated data may only be generalizable to the data that meet the conditions under which the data are generated. In real life situations, sometimes, not all examinees answer all items (e.g. the data used in this study). Also, guessing is assumed to occur in many testing situations. With easy

items, guessing is not as frequent as with difficult items. With very difficult items, more guessing may occur, which may influence the performances of the MH and LR statistics. Guessing can be built into simulated data, but the guessing parameter is usually held constant for all the items under study. In reality, different levels of guessing may occur depending on the background and knowledge of examinees. Moreover, not every examinee guesses.

Some researchers have employed real examinee data to study the performance of the MH procedure (e.g. Clauser, et al., 1991). Brown (1992) has used real examinee data to compare the MH and LR procedures. However, they have conducted no replications or inadequate numbers of replications due to limited data (5 at most in Brown's study).

Several researchers have reported that the LR procedure was more powerful than the MH procedure for detecting nonuniform DIF and as powerful in detecting uniform DIF. However, no assurance can be made that the simulated data used in their study compare to a real life situation. The validity of their claims is worth testing by an empirical study using real examinee data.

Sample size affects the power of the LR and MH procedures. In small samples the test statistic may not be a valid indicator of the presence of DIF. With the MH procedure, small samples may influence the stability of the estimate of the odds ratio because it is weighted by the number of examinees in each score interval. Both the MH and LR procedures use a chi-square test of significance. It is known that the chi-square test statistic is inflated by large samples. It was found that the power of the MH and LR statistics improved with larger samples, but there was no agreement on how large or how

small the sample size should be in order to obtain consistent MH and LR statistics. So far, very little work has been done on the consistency of the MH procedure and especially the LR procedure across different sample size when using real examinee data. To provide information regarding optimal sample size needed when using the MH and LR procedures, a study of their performance under different sample sizes when using real examinee data is appropriate.

Ackerman (1992) noted that the main cause of item bias is the misspecification of the latent ability space, where items that measure multiple abilities are scored as though they are measuring a single ability. When a test is intentionally multidimensional, DIF analyses using total test score as the matching variable are less likely to be successful. The dimensionality of the matching criterion has always been an important issue in detecting DIF. The MH and LR procedures incorporate the notion of comparing the item performance of members of different groups who are in some sense comparable. The typical DIF detection approach to ensure comparability of group members is to use the total test score as a control variable because it is commonly assumed that the total test score is the most practical and appropriate criterion against which group performance on individual items within a test can be compared (Clauser, et al., 1991). Total test score is used as the predictor in the LR model and as the criterion for grouping examinees in the MH procedure (Swaminathan & Rogers, 1990). However, tests appearing to be unidimensional may sometimes include subtests or items that measure more than one skill or another skill. Clauser, et al. (1991) indicated that systematically changing the grouping of test items analyzed resulted in some items showing substantial changes in

their MH statistics. Therefore, another concern for study is whether the choice of criterion for measuring ability has any effect on the stability of the MH and LR procedures in the identification of DIF.

Generally speaking, the MH and LR procedures are simple and practical and can be theoretically justified. Unfortunately, limited information is available regarding the performance of the two procedures, especially when real examinee data are used. To provide information on their performance with real data, further research needs to be conducted on a large and real examinee dataset so that random sampling and a large number of replications can be achieved.

Based on the above, the purpose of this study was to examine the performance of the LR and MH procedures and their agreement in the identification of DIF under different conditions when using real data. This study is especially intended to provide information on how the MH and LR procedures behave with real examinee data. Three specific research questions are addressed in this study:

1. How consistently do the MH and LR procedures identify items as showing DIF across sample size?

2. Does the choice of the criterion for measuring ability have any effect on the performance of each of the two procedures?

3. Do the MH and LR procedures show agreement in the identification of DIF?

CHAPTER III

METHODOLOGY

In this chapter the methodology used to assess the performance of the MH and LR procedures with real data is presented. The chapter is divided into three sections: data description, procedures, and data analysis.

Data Description

The ACT Assessment Program is a comprehensive data collection, processing, and reporting system. A major part of the ACT Assessment Program is a test battery which includes tests in English, Mathematics, Reading, and Science Reasoning. These tests measure skills and abilities highly related to high school course work and are used as predictors of success in colleges and universities. For this study, data from the 1989 administration of the ACT Reading Test (Form 39B) were used. The ACT Reading Test (Form 39B) is a 40-item test which consists of four reading passages. Based on each passage, there are 10 multiple choice questions. These questions require not only reading comprehension skills, but also the ability to draw inferences and conclusions, to examine the interrelationships and importance of ideas in a passage, and to recognize a writer's style and mode of reasoning. Passage 1 (P1) is a literature passage. Passage 2 (P2) describes light and colour in nature. Passage 3 (P3) is about theatrical production in ancient Greek. The last passage (P4) is related to psychology and is a discussion of creativity. Items based on passages 1 and 3 form the Reading Literature/Arts subtest

(Sub1). Items based on passages 2 and 4 constitute the Reading Science/Social Studies subtest (Sub2).

The available dataset included the item responses of 244,119 examinees of the 1989 administration of the ACT Assessment (Form 39B). The examinees were college-bound high school seniors mostly from different states of the U.S. In this study, the item responses of 183,008 Caucasian examinees (75,752 males and 107,256 females) who took the Reading Test were used as the database. The descriptive statistics for total test and subtests based on 183,008 Caucasian examinees are presented in Table 2.

Table 2: Descriptive Statistics: ACT Reading Test

	Total (40 items)			Sub1 (20 items)			Sub2 (20 items)		
	Mean	SD	KR-20	Mean	SD	KR-20	Mean	SD	KR-20
Male	22.33	6.43	.82	11.85	3.61	.71	10.48	3.63	.72
Female	22.64	6.13	.80	12.32	3.46	.69	10.32	3.54	.71

From the means and standard deviations, one can see that females and males have approximately equal ability distribution. The reliability of the total test (KR-20) is moderately high, but decreased for subtests. This is understandable since there are fewer items in the subtests. Overall, the reliability of both total test and subtests was slightly lower for females than for males. The correlation between the two subtests is .55, which is very low. Only about 30% of variance was in common in each subtest.

The descriptive statistics of the four reading passages are presented in Table 3. The means of the females are slightly higher in Passages 1 and 4 but slightly lower in

Passages 2 and 3 than those of the males. However, the difference is very small.

Table 3: Descriptive Statistics of the Four Reading Passages

	P 1 (Literature)		P 2 (Science)		P 3 (Arts)		P 4 (Social Studies)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Male	6.67	2.05	6.15	1.78	5.18	2.19	4.33	2.65
Female	7.26	1.95	5.77	1.66	5.06	2.13	4.55	2.75

The correlations among the four reading passages are reported in Table 4. They are even lower than the correlation between the two subtests. The low correlations are very surprising, which seem to suggest that the four passages share very little in common and may be measuring four different reading abilities.

Table 4: Correlations Among the Four Reading Passages

	P 1	P 2	P 3
P 2	.36		
P 3	.43	.41	
P 4	.32	.27	.42

Procedures

Sampling

Using Holland and Thayer's terminology, female examinees were classified as focal group members (subgroup of interest) and males were classified as reference group members (standard to compare performance of the focal group). Grouping in this study

is arbitrary. Comparison groups were created by randomly selecting cases by gender across four levels of sample size with an equal number of examinees in both reference and focal groups: 100/100, 250/250, 500/500, and 1000/1000. Thirty gender comparison groups were drawn for each level of sample size by random sampling with replacement. These sample sizes were selected for investigation because they covered a broad range and were reported in the literature to reflect the kinds of sample sizes available to practitioners.

Computation of the DIF Statistics

Computation of the MH and LR statistics was performed using computer programs developed by Ackerman (1987) and Spray (1991), respectively. With the LOGREG program (Spray, 1991), the chi-square statistics for testing the hypotheses of no uniform or nonuniform DIF were calculated separately, each with one degree of freedom. With the Program MANTEL (Ackerman, 1987), the MH chi-square and MH-Z were computed. As was noted before, MH-Z is the natural log of α_{MH} divided by (-1.7), and it is equivalent to Δ_{MH} divided by 4.

In order to examine the effectiveness of the MH and LR statistics in detecting DIF when the criterion for measuring ability is varied, the two procedures were first applied to the Reading Test using total test score as the criterion. Then, the statistics were calculated for items in each of the two subtests using subtest score as the criterion. Finally, the statistics were recalculated for items based on each of the four passages using passage score as the criterion. Using the passage score as the criterion was not

originally planned. A surprisingly large number of items were detected when subtest score was used as the criterion. Based on the low correlations among the four passages, a decision was made to use passage score as the criterion to see if some DIF items would disappear. In order to minimize the impact of chance variability, 30 replications were conducted for each combination of sample size and criterion variable.

Holland and Thayer's (1986) suggestion of purifying the matching criterion for calculating the MH statistics was not addressed in this study since it is very tedious and was deemed unnecessary to conduct this step with many replications.

Data Analysis

The mean and standard deviation of the MH-Z values and frequency of DIF identification by the MH and LR chi-square statistics over 30 replications at two significance levels (.01 and .05) were computed for each item for each combination of sample size and criterion variable. The agreement between the MH and LR procedures and the consistency of the two procedures across sample size and criterion were examined according to the number of times an item was identified. The MH-Z was used to determine the direction (i.e. favouring males or females) and size of difference in item difficulty. A positive value of MH-Z indicated that an item was easier for females, while a negative value meant that the item favoured males. Unlike DIF studies where simulated data are used, there is no means in this study for pre-identifying DIF items. Therefore, rules had to be specified in interpreting the results of DIF analyses. Three rules were developed to classify items according to their DIF statistics based on mean MH-Z values

over 30 replications at sample size of 1,000:

1. No Serious DIF $|MH\text{-}Z| < .20$
2. Possible DIF $.20 \leq |MH\text{-}Z| < .25$
3. Definite DIF $|MH\text{-}Z| \geq .25$

That is, items with an absolute mean MH-Z less than .20 ($|\Delta_{MH}| < .8$) were considered as free of DIF or not having serious DIF; items with an absolute mean MH-Z value equal to or greater than .20 but less than .25 ($.8 \leq |\Delta_{MH}| < 1$) were classified as possible DIF items; items with an absolute mean MH-Z value equal to or greater than .25 ($|\Delta_{MH}| \geq 1$) were identified as definite DIF items. These rules are less stringent than those used by ETS. The rules used by ETS are based on one replication only. The reason for employing less rigorous standards is that the mean of MH-Z values for each item over 30 replications instead of the MH-Z values for one replication only was used to classify items. In using these criteria the sample size of 1,000 was chosen as the standard because it represents the largest sample size in this study and the MH-Z values should be relatively stable at this sample size.

Since the MH procedure is not designed to detect nonuniform DIF, the above-specified rules can only be applied to classify items with uniform DIF. By referring to the above rules, it was decided that items which had been detected at least 20 times out of 30 replications by the LR nonuniform statistic at the .01 significance level were considered as possible DIF items, and items which had been detected at least 25 times out of 30

replications by the LR nonuniform statistic at .01 significance level were considered as definite DIF items.

CHAPTER IV

RESULTS

In this chapter, the results of this study are presented. Included are seven sections: DIF classification, the influence of the criterion variable, sample size effect, nonuniform DIF, the agreement between the MH and LR procedures, and Type I error rate.

DIF Classification

The means and standard deviations of MH-Z values over 30 replications for items identified as having Possible or Definite DIF are reported in Table 5. The means and standard deviations of MH-Z values over 30 replications for all the forty items are reported in Table 9 (APPENDIX A). Using the total test score as the criterion for measuring ability, four items (Items 1, 2, 5, 19) were identified as Definite DIF and three items (Items 17, 18, 20) fell into the category of Possible DIF (See Table 5) according to the prespecified classification standards based on the MH-Z. The remaining items were considered as free of DIF, or the DIF effect was not serious.

The DIF detection by the MH-Z statistic using subtest score as the criterion variable indicated some change of results. Altogether twelve items were identified as showing DIF; six items were in common with the items identified in the total test and six new DIF items appeared. Item 18 was no longer a DIF item. Item 2 was now identified as a Possible DIF item. In Subtest 1 (Reading Literature/Arts), five items were identified with Items 1, 5, and 23 being Definite DIF. Items 2 and 26 were Possible DIF. Seven items were flagged in

Table 5 The Means and Standard Deviations of MH-Z Over 30 Replications for Items Identified as Having Possible or Definite DIF Using Total Test, Subtest or Passage Score as the Criterion

		Total Test		Subtest		Passage	
Items	Sample Size	Mean	SD	Mean	SD	Mean	SD
1	1,000	**.32**	.11	**.28**	.09	.17	.11
Total**	500	**.31**	.16	**.28**	.14	.17	.11
Sub1**	250	**.33**	.24	**.27**	.26	.18	.27
P1	100	**.46**	.31	**.40**	.27	.26	.27
2	1,000	**.26**	.07	**.21**	.07	.07	.09
Total**	500	**.24**	.09	**.20**	.08	.07	.09
Sub1*	250	**.26**	.13	**.23**	.13	.11	.15
P1	100	**.26**	.21	**.21**	.23	.05	.26
5	1,000	**.43**	.08	**.41**	.06	**.30**	.09
Total**	500	**.45**	.11	**.43**	.09	**.30**	.09
Sub1**	250	**.44**	.14	**.42**	.15	**.30**	.16
P1**	100	**.44**	.23	**.43**	.23	**.29**	.26
13	1,000	.11	.11	.19	.08	**.32**	.09
Total	500	.12	.11	.17	.11	**.31**	.12
Sub2	250	.12	.13	.17	.16	**.30**	.15
P2**	100	.14	.33	.23	.29	**.37**	.26
14	1,000	.08	.07	.12	.07	**.23**	.06
Total	500	.09	.07	.12	.07	**.23**	.09
Sub2	250	.10	.12	.15	.12	**.26**	.12
P2*	100	.05	.22	.11	.23	**.26**	.22
17	1,000	**-.24**	.05	**-.21**	.05	-.11	.05
Total*	500	**-.23**	.06	**-.22**	.06	-.13	.07
Sub2*	250	**-.25**	.13	**-.22**	.13	-.13	.15
P2	100	**-.26**	.15	**-.26**	.13	-.15	.13
18	1,000	**-.20**	.05	-.16	.05	-.04	.05
Total*	500	**-.19**	.10	-.15	.10	-.02	.10
Sub2	250	**-.18**	.14	-.15	.13	-.02	.14
P2	100	**-.22**	.18	-.19	.17	-.06	.17
19	1,000	**-.36**	.06	**-.32**	.05	**-.21**	.06
Total**	500	**-.35**	.10	**-.31**	.09	**-.19**	.09
Sub2**	250	**-.35**	.10	**-.32**	.11	**-.21**	.12
P2*	100	**-.34**	.20	**-.33**	.20	**-.20**	.22

Table 5 (Continued)

		Total Test		Subtest		Passage	
Items	Sample Size	Mean	SD	Mean	SD	Mean	SD
20	1,000	**-.23**	.08	**-.21**	.08	-.10	.09
Total*	500	**-.26**	.10	**-.23**	.12	-.12	.13
Sub2*	250	**-.24**	.15	**-.22**	.14	-.13	.15
P2	100	**-.28**	.32	**-.29**	.36	-.14	.37
23	1,000	-.18	.07	**-.26**	.07	-.13	.07
Total	500	-.16	.09	**-.22**	.09	-.09	.09
Sub1**	250	-.20	.10	**-.28**	.10	-.16	.11
P3	100	-.19	.19	**-.29**	.19	-.11	.16
26	1,000	-.14	.06	**-.20**	.06	-.09	.06
Total	500	-.13	.09	**-.19**	.09	-.08	.10
Sub1*	250	-.15	.12	**-.22**	.12	-.10	.12
P3	100	-.24	.20	**-.25**	.21	-.15	.21
34	1,000	.10	.07	**.20**	.08	.06	.09
Total	500	.14	.09	**.23**	.10	.09	.10
Sub2*	250	.15	.13	**.22**	.14	.09	.16
P4	100	.19	.23	**.26**	.24	.13	.28
35	1,000	.16	.05	**.23**	.06	.14	.06
Total	500	.13	.09	**.19**	.09	.10	.09
Sub2*	250	.12	.12	**.18**	.12	.10	.13
P4	100	.06	.23	**.14**	.21	.06	.22
36	1,000	.17	.07	**.27**	.08	.12	.08
Total	500	.17	.08	**.27**	.09	.11	.09
Sub2**	250	.16	.11	**.25**	.13	.11	.14
P4	100	.15	.22	**.26**	.23	.08	.24
37	1,000	.16	.06	**.24**	.06	.11	.07
Total	500	.16	.11	**.21**	.10	.08	.10
Sub2*	250	.19	.12	**.26**	.13	.14	.11
P4	100	.18	.22	**.26**	.22	.09	.19

Note: * Possible DIF, ** Definite DIF. For example, Item 2 was identified as Definite DIF using total test score as the criterion, Possible DIF using subtest score as the criterion, and no DIF using passage score as the criterion.

Subtest 2 (Reading Science/Social Studies). Among them, Items 19 and 36 were Definite DIF. Items 17, 20, 34, 35, and 37 were Possible DIF (See Table 5).

Using the passage score as the criterion, the number of items identified was reduced substantially. Only four items were identified, two of which were new DIF items (Item 13 and Item 14). Items 5 and 13 were Definite DIF. Items 14 and 19 were Possible DIF (See Table 5). Items 5 and 19 were identified under all three criteria. However, Item 19 fell into a different category of DIF (Possible DIF) when passage score was used.

Influence of the Criterion Variable

The influence of the criterion variable is very obvious. In addition to different numbers of items and different items identified when using different criteria, it should also be noted that when the criterion variable was subtest score, the absolute means of MH-Z were generally lower for items in Passages 1 and 2 and higher for items in Passages 3 and 4 than when the criterion was total test score (See Table 5 and Table 9). This is likely due to the fact that for each subtest the mean of the females or males is slightly higher on one passage but lower on the other (See Table 3 on page 43). The means of the females are slightly higher on Passages 1 and 4 but slightly lower on Passages 2 and 3 in this dataset. In examining Table 5 and Table 9 (APPENDIX A), the direction of DIF was such that it favoured the group scoring higher on the passage. DIF items in Passages 1 and 4 favoured females, and in Passages 2 and 3 favoured males. These were the results observed when the items were examined with the criteria of total test score and subtest score.

The low correlations (See Table 4 on page 43) among the four passages seem to suggest that the four passages are measuring four different reading abilities. As a result, using total test score and especially subtest score as the criteria for measuring ability when investigating DIF resulted in too many DIF items detected by the MH and LR procedures. Also by examining the raw data, it was found that a number of examinees did not answer the last five to ten items, which implicated an additional speed factor underlying the examinees' performance.

When the criterion shifted from total test score to passage score, an even larger decrease of mean MH-Z values occurred for most of the items in passages 1 and 2 except for mean MH-Z values of Items 11, 13, and 14 which increased substantially. However, there was a general trend of decrease of the mean MH-Z values for items in passages 3 and 4 (See Table 5 and Table 9). No general trend of the direction of DIF was observed. An interesting phenomenon is that three of the four items detected using passage score as the criterion were in Passage 2. Using total and subtest score as the criteria, DIF items identified in Passage 2 all favoured males. However, when passage score was used, the MH-Z picked up two items with opposite effect, two (Item 13 and Item 14) of the three items identified in Passage 2 favoured females. This DIF effect against males would have gone undetected without using passage score as a criterion. These two items were not identified in the total and subtest. Since items based on each passage are associated to a specific content area--Literature, Science, Arts, or Social Studies, there seemed to exist a passage or context effect. However, It was very hard to separate the passage effect from the test length effect since there are only ten items in

each passage. Nevertheless, the effect of test length is interesting since more items were identified in subtest than in total test and passage. The numbers of items identified in total test, subtest, and passage were 7, 12, and 4, respectively.

When the criterion shifted from subtest score to passage score, there was a general decrease of MH-Z values except for a few items (e.g. Item 13 and Item 14). The number of items identified was greatly reduced. The direction of DIF is very consistent between using total test score and using subtest score as the criteria, but inconsistent between using total test score and using passage score, and between using subtest score and using passage score as the criteria. The change of the results from using subtest score to using passage score as the criterion is similar to that observed from using total test score to using passage score as the criterion.

By further examining Table 5 and Table 9, it was noticed that the direction of DIF was more extreme in total and subtest but became balanced in passage. Take Passage 3, for example; when the total test was used as the criterion variable, the ratio between items that favoured males and items that favoured females was 7:3. When the criterion was subtest, the ratio was 9:1. But when the criterion was passage score, the ratio became 5:5. That means there were equal number of items favouring males or females when the passage score was used as the criterion variable. Similar results were observed for items in the other three passages (See Table 9).

As for DIF items, three DIF items favoured females in Passage 1 and four DIF items favoured males in Passage 2 when the total test score was used as the criterion. When the criterion changed to subtest score, three items in Passage 1 favoured females; three

items in Passage 2 favoured males; two items in Passage 3 favoured males; four items in Passage 4 favoured females. When the criterion was passage score, one item in Passage 1 and two items in Passage 2 favoured females; one item in Passage 2 favoured males.

In Table 6, the frequency of DIF identification by the MH chi-square statistics and LR uniform (LR(U)) and nonuniform (LR(N)) chi-square statistics at the .01 significance levels for items identified as Possible or Definite DIF is presented. The frequency of DIF identification at the .01 and .05 significance level for all the forty items are presented in Table 10 (APPENDIX B) and Table 11 (APPENDIX C), respectively. Identifying DIF at the .01 significance level was the main concern of this study. The DIF detection results at the .05 significance level were reported for reference and comparison. Similar results to those based on MH-Z were found for the detection rates of the MH and LR Uniform chi-square statistics at the .01 significance levels. As with the size of the MH-Z, when the criterion changed from total test to subtest, the frequency of DIF detection decreased for items in Passages 1 and 2 and increased for items in Passages 3 and 4. When passage score was used as the criterion, the frequency of DIF identification was close to zero for most of the items (See Table 6, Table 10, and Table 11). As noted above, two items behaved differently when the criterion changed from total test to passage and from subtest to passage. They are Items 13 and 14. Their frequencies increased from close to zero to 23 and 26 respectively for sample size of 1,000 at the .01 significance level.

In order to see if these results were caused by the multidimensionality of the criterion for measuring ability, an analysis of the dimensionality of the total and subtest was

Table 6 The Frequency of DIF Identifications over 30 Replications by the MH and LR Chi-Square Statistics at the .01 Significance Level for Items Identified as Having Possible or Definite DIF

Items	Sample Size	Total Test			Subtest			Passage		
		MH	LR(U)	LR(N)	MH	LR(U)	LR(N)	MH	LR(U)	LR(N)
1	1,000	**25**	**26**	1	**16**	**17**	0	2	3	0
Total**	500	**13**	**15**	0	**6**	**7**	0	2	3	0
Sub1**	250	**5**	**5**	0	**4**	**5**	0	4	5	0
P1	100	**1**	**1**	1	**0**	**2**	0	0	1	0
2	1,000	**26**	**28**	1	**19**	**19**	3	2	2	1
Total**	500	**13**	**14**	2	**9**	**9**	1	2	2	1
Sub1*	250	**9**	**10**	1	**6**	**8**	1	1	2	0
P1	100	**1**	**3**	0	**1**	**2**	0	1	2	0
5	1,000	**29**	**30**	4	**30**	**30**	2	**24**	**24**	0
Total**	500	**28**	**29**	3	**29**	**29**	1	**24**	**24**	0
Sub1**	250	**22**	**25**	2	**21**	**22**	1	**9**	**10**	0
P1**	100	**6**	**16**	0	**11**	**14**	0	**1**	**2**	0
13	1,000	2	2	0	7	7	1	**23**	**23**	3
Total	500	0	0	0	1	2	0	**9**	**10**	1
Sub2	250	0	0	1	0	1	0	**1**	**2**	0
P2**	100	0	0	0	0	0	0	**0**	**2**	0
14	1,000	2	2	0	6	6	0	**26**	**26**	2
Total	500	1	1	1	2	3	1	**14**	**14**	1
Sub2	250	1	1	0	3	3	0	**9**	**9**	0
P2*	100	0	0	0	0	0	0	**2**	**2**	0
17	1,000	**30**	**30**	3	**28**	**29**	1	6	7	0
Total*	500	**21**	**22**	1	**15**	**19**	2	1	4	1
Sub2*	250	**11**	**12**	0	**7**	**10**	0	2	4	1
P2	100	**0**	**4**	0	**1**	**2**	0	0	0	0
18	1,000	**25**	**25**	2	17	19	2	0	1	0
Total*	500	**14**	**13**	0	10	11	0	0	1	0
Sub2	250	**6**	**7**	0	4	6	0	0	0	0
P2	100	**0**	**1**	0	0	0	1	0	0	1
19	1,000	**30**	**30**	1	**30**	**30**	3	**24**	**26**	1
Total**	500	**26**	**29**	1	**25**	**27**	2	**10**	**13**	2
Sub2**	250	**18**	**19**	0	**15**	**18**	0	**2**	**4**	0
P2*	100	**4**	**7**	1	**5**	**6**	0	**2**	**3**	0

Table 6 (Continued)

Items	Sample Size	Total Test			Subtest			Passage		
		MH	LR(U)	LR(N)	MH	LR(U)	LR(N)	MH	LR(U)	LR(N)
20	1,000	**18**	**18**	1	**12**	**17**	2	4	7	0
Total*	500	**12**	**15**	0	**8**	**12**	1	1	2	0
Sub2*	250	**1**	**2**	1	**1**	**2**	2	0	1	2
P2	100	**1**	**3**	0	**2**	**3**	0	1	1	1
23	1,000	19	19	0	**28**	**29**	0	8	8	0
Total	500	5	5	1	**11**	**16**	1	1	1	1
Sub1**	250	3	6	0	**9**	**10**	0	3	3	1
P3	100	0	0	0	**2**	**3**	0	0	1	0
26	1,000	12	12	1	**19**	**22**	2	2	5	0
Total	500	4	7	1	**10**	**11**	0	2	4	0
Sub1*	250	2	2	0	**5**	**5**	1	1	1	2
P3	100	0	1	0	**0**	**1**	0	0	0	0
34	1,000	6	7	1	**21**	**22**	2	4	4	0
Total	500	4	4	0	**13**	**13**	0	0	1	1
Sub2*	250	4	3	0	**4**	**4**	1	2	2	0
P4	100	1	1	1	**2**	**3**	0	2	2	2
35	1,000	14	13	1	**26**	**27**	1	10	9	0
Total	500	5	5	0	**11**	**12**	0	2	1	0
Sub2*	250	2	1	0	**4**	**4**	0	0	0	0
P4	100	1	1	0	**1**	**2**	0	0	0	0
36	1,000	16	14	0	**27**	**26**	1	6	7	0
Total	500	3	3	1	**19**	**19**	2	1	1	0
Sub2**	250	0	1	0	**4**	**7**	1	0	0	0
P4	100	1	1	0	1	**2**	1	1	1	0
37	1,000	12	12	1	**25**	**27**	3	5	4	0
Total	500	8	6	1	**9**	**10**	1	1	1	2
Sub2*	250	2	3	0	**4**	**6**	1	1	1	0
P4	100	1	2	1	**2**	**4**	0	0	0	1

conducted using program NOHARM II on a random sample of 1,000 examinees (including both males and females). Passages were not analyzed since there were too few items (10) in each passage. The dimensionality analyses were also performed separately for males and females. The results obtained were quite similar to those based on the combined sample of males and females.

The results of the chi-square test of the fit of the model with 740 degrees of freedom indicated that the total reading test is not unidimensional. Following this, a two-factor solution was requested. Although the resulting chi-square test statistic was still greater than the critical value, the other evidence such as residual matrix and percent of z-values greater than 1.96 (5%) suggested that the model fit the data. The factor loadings (See Table 7) obtained from NOHARM II show that the test contains a dominant factor with one additional factor. Items 1 to 30 seemed to load on the first factor. Items 31 to 40 seemed to load on both the first and the second factor but more heavily on the second factor. The dominant factor appeared to be the general reading ability while the smaller factor appeared to be a speed factor. The same data were again analyzed using three-, four-, and even five-factor models, but there was no clear indication of dimensions. These results were not reported for this study.

For Subtest 1, the chi-square test statistic with 170 degrees of freedom for a unidimensional model is 213, which is slightly greater than the critical value of 207. The percent of z-values greater than 1.96 is .07. Although the hypothesis of unidimensionality was rejected, a two-factor solution did not produce a clear indication of two factors (See Table 8), which suggested that Subtest 1 was approximately unidimensional. As for

Subtest 2, there clearly existed two factors with each factor related to each passage (See Table 7).

This evidence seems to indicate the multidimensionality of the criterion is the cause of the results of this study. Another possible reason might be test length. However, in some way, test length might not be the issue since there were more items identified in subtest than in total test.

Sample Size Effect

Tables 6 and Table 10 show that sample size had a strong influence on the performance of the MH and LR chi-square statistics. As expected, the power of the MH and LR chi-square statistics improved substantially as sample size increased. At sample size of 1,000, the frequency of DIF detection is very high even when the mean effect size (MH-Z) is relatively small (e.g. Item 10 and some other items). At sample size of 100, the detection rates for both the MH and LR statistics at .01 significance level were close to zero for most of the items even when the mean effect size was large. The exception was Item 5. It should be noted that Item 5 had the largest mean effect size. The means of the MH-Z for Item 5 are .43, .41, and .30 at sample size of 1,000 when the criteria were total test, subtest, and passage, respectively. The influence of sample size on this item is not as obvious as for other items. The difference between the detection rates at .01 level and at .05 level for Item 5 are also not as large as that for other items. However, the detection rates are still very low at sample size of 100. With a mean effect size of .44 in total test for Item 5 but still low detection rates (6 times out 30 replications for the MH procedure

Table 7 Factor Loadings of Items in the Total Test and Subtests

	Total Test		Subtest 1		Subtest 2	
	F1	F 2	F1	F 2	F1	F2
1	.50	.00	.51	.00	.47	.00
2	.56	.04	.54	-.16	.59	.04
3	.52	-.02	.52	-.24	.57	.03
4	.40	-.02	.41	-.19	.37	-.01
5	.51	.04	.58	-.40	.45	.09
6	.28	.12	.26	.07	.52	.08
7	.50	.07	.53	-.14	.24	.10
8	.37	.04	.39	-.22	.44	.14
9	.30	-.01	.31	-.16	.37	.16
10	.35	.01	.33	-.12	.14	-.07
11	.49	-.12	.60	.05	.23	.64
12	.45	-.03	.29	.12	.16	.71
13	.38	-.01	.51	.16	.23	.65
14	.32	-.06	.55	.17	.24	.70
15	.39	.02	.52	.09	.02	.75
16	.41	.03	.36	.11	.13	.76
17	.22	.08	.38	.22	.15	.62
18	.41	.07	.49	.18	.17	.27
19	.38	.09	.33	.23	.15	.37
20	.17	-.13	.37	.29	.19	.53
21	.55	.10				
22	.26	.07				
23	.50	.09				
24	.51	.15				
25	.53	.08				
26	.35	.18				
27	.35	.15				
28	.48	.23				
29	.30	.20				
30	.35	.25				
31	.39	.56				
32	.27	.68				
33	.40	.57				
34	.46	.60				
35	.17	.73				
36	.29	.72				
37	.31	.55				
38	.18	.24				
39	.17	.35				
40	.23	.50				

and 16 times out of 30 replications for the LR procedure at the .01 significance level) at sample size of 100, it definitely indicates that a sample size of 100 is not approximate.

As expected, the means of the MH-Z were relatively consistent across sample size. However, the standard deviation over 30 replications of the MH-Z for each item increased as the sample size decreased. For items identified as having DIF, the standard deviations of MH-Z ranged from .05 to .11 at sample size of 1,000 and from .13 to .32 at sample size of 100. Similar variability was observed for items not identified as having DIF (See Table 5 and Table 9). This is expected since the MH-Z values should be more stable for large samples than for small samples.

There are a number of items that look as though they are biased based on the frequency of DIF identification (See Table 10 and Table 11). For example, Item 16 and Item 25 were not classified as DIF items according to the rules set in this study. However, their detection rates (24 times for Item 16 for total test and 21 times for Item 25 for Subtest 1 at the .01 level) are higher than some of the DIF items (e.g. Item 1 and Item 20). In examining Table 5 and Table 9, it was found that these two items had smaller standard deviations for MH-Z (.06 at sample size of 1,000) comparing to some DIF items. Also, some definite DIF items (e.g. Item 1) had lower detection rates than some possible DIF items (e.g. Item 17). Item 17 also had a much lower standard deviation for MH-Z than Item 1. The mean MH-Z of Item 1 is large (.32 and .28 at sample size of 1,000 in Total Test and Subtest, respectively). However, its detection rates are not as large as expected, especially when the criterion was subtest score. From the standard deviations of MH-Z for Item 1 (.11 and .09 at sample size of 1,000 in Total Test and Subtest,

respectively), one can see that there might be some very small MH-Z values which reduced the detection rates of this item. This raised question against using small samples, especially sample sizes below 500, because the standard deviations of the MH-Z are very high and the detection rates are very low.

Nonuniform DIF

According to the prespecified rules, no definite or possible nonuniform DIF items were identified. When the total test score was used, item 29 produced the highest frequency of nonuniform DIF detection at sample size of 1,000 and at the .05 significance level (16 times out of 30 replications). When the criterion was subtest score, item 8 had the highest detection rate (18 times out of 30 replications at sample size of 1,000 and at the .05 level). Item 10 had the highest nonuniform DIF detection frequency when the passage score was used (15 times at sample size of 1,000 and at the .05 level). The frequencies of nonuniform DIF detection for other items were all close to zero (≤5 times) except for a few items (5) which have detection rates close to 10 at the .05 significance level. As expected, the MH procedure did not detect these items as frequently (See Table 11). As expected, the means of MH-Z for Items 8, 10, and 29 are very low (See Table 9).

The Agreement Between the MH and LR Procedures

A comparison of the frequency of DIF identification by the MH and LR Uniform procedures shows that, in general, the agreement between the two procedures is very high in the identification of uniform DIF across sample size and under different criteria

with the LR procedure producing slightly higher detection rates. However, as sample size decreased, the discrepancy between the two procedures became larger (e.g. Item 5).

Type I Error Rate

To provide additional information, the false-positive rates of the MH and LR procedures were computed based on items which have absolute mean MH-Z values of .05 or smaller at each sample size for each criterion. Thus the false-positive rate was the proportion of times these items were identified as having DIF. The results are shown in Table 8. Since real data were used and the true conditions of each item were unknown, the values in Table 8 are not truly false-positive rates because they were based on items with MH-Z values smaller than .05. These rejection rates could be greater than the nominal level of significance because of the amount of DIF that is present in the item. The purpose of calculating these values was to get a rough idea of the false rejection rates of the MH and LR procedures.

Overall, the false-positive rate increased as sample size increased. At the .01 significance level, the MH and LR procedures produced an average false-positive rate of .007 and .009, respectively, across three criterion variables. With a significance level of .05, the average false-positive rates produced by the MH and LR procedures under the three criteria are .041 and .049, respectively. The false-positive rates of both procedures were within the expected range with the LR procedure producing a slightly higher rate, which is more obvious at the .05 significance level. The difference may account for the slightly higher detection rate for the LR procedure.

Table 8 Mean False Positive Rates for the MH and LR Uniform Procedures

		.01		.05	
	Sample	MH	LR	MH	LR
T	1,000	.025	.025	.072	.078
o	500	.006	.009	.024	.078
t	250	.008	.010	.042	.047
a	100	.003	.007	.017	.050
l					
S	1,000	.017	.022	.083	.089
u	500	.017	.022	.072	.072
b	250	.000	.007	.067	.060
1	100	.000	.000	.011	.011
S	1,000	.000	.000	.033	.044
u	500	.000	.000	.033	.033
b	250	.000	.000	.011	.056
2	100	.017	.017	.083	.083
	1,000	.000	.000	.044	.044
P	500	.000	.000	.044	.044
1	250	.000	.011	.044	.033
	100	.008	.025	.033	.083
	1,000	.008	.017	.083	.083
P	500	.000	.008	.042	.050
2	250	.000	.008	.025	.025
	100	.000	.000	.022	.033
	1,000	.050	.000	.067	.050
P	500	.017	.000	.022	.033
3	250	.020	.020	.053	.073
	100	.000	.000	.020	.033
	1,000	.000	.000	.000	.000
P	500	.000	.000	.011	.033
4	250	.000	.017	.022	.033
	100	.000	.000	.033	.033

CHAPTER V

DISCUSSION

Consistent with Swaminathan and Rogers (1990) and other researchers, the findings in this study indicate that the power of the MH and LR chi-square statistics increased as the sample size increased. One should not be surprised since any statistic will be more powerful as the sample becomes larger. The mean MH-Z values are relatively consistent across sample size. However, the MH-Z statistic is less subject to sampling fluctuations as the sample size increased.

The implications for practitioners are clear. The results of the MH and LR procedures are questionable at small sample sizes because of the low detection rates even for large effect size. It seems that a sample size less than 500 is too small to yield reliable MH and LR statistics unless the effect size is very large such as for Item 5. This is also evidenced by the large standard deviations of the MH-Z observed for sample sizes under 500. Even items showing large amounts of DIF may go undetected when the sample size is small. A sample size of 500 or larger seems to produce relatively more stable MH and LR indices. For items with large MH-Z values (e.g. Item 5), the difference between the detection rates at sample sizes of 500 and 1000 is very small.

In this study, the performance of the MH and LR procedures was found to be affected by the criterion variable. Only two items were in common across the three criteria. When using the total test score as the criterion for measuring ability, seven items were identified based on the mean MH-Z values. When using subtest score as the criterion variable, the

means of MH-Z values and the frequency of DIF detection by the MH and LR chi-square indices dropped for some items and increased for other items. When items were reanalyzed in the context of the two subtests, six new DIF items were identified. One item no longer had DIF and another item fell into a different category of DIF classification. When passage score was used as the criterion, the frequency of DIF detection and the means of MH-Z showed a general trend of decreasing except for a few items. The number of items identified was greatly reduced. However, two new DIF items appeared. It appears that the context in which items were studied might influence the results. Some of the findings are consistent with Clauser, et al. (1991) in which the choice of criterion, total test score versus subtest score, had a substantial influence on the performance of the MH statistics. However, the findings in this study are different from Tian, Pang, and Boss (1994) in which the DIF detection by either the MH procedure or the LR procedure was not affected by the changing of criterion variable for the 1989 ACT English Test dataset.

Surprisingly high numbers of items were identified in this study especially when the subtest score was used as the criterion. The multidimensionality of the test is probably the cause of the results in this study. Obviously, the Reading Test and even the two subtests were composed of different types of reading items. Also the results of dimensionality analyses by using Program NOHARM II showed that these tests were not absolutely unidimensional. Furthermore, test items are related to four different scenarios and the intercorrelations among the four passages are very low. Ackerman (1992) noted that the main cause of item bias is the misspecification of the latent ability space. If two

groups of examinees have different underlying multidimensional ability distributions and the test items are capable of discriminating among levels of abilities on these multiple dimensions, then any unidimensional scoring scheme has the potential to produce item bias. The results of this study seemed to have provided evidence to support Ackerman's (1992) argument. The results of Tian et al. (1994) implied that two English subtests (Usage/Mechanics and Rhetorical Skills) seemed to be measuring one general English ability. As a result, when the internal criterion was varied, no meaningful changes were observed on the performance of the two procedures. It was very interesting to see that some DIF items identified in the total reading test and the two subtests disappeared when the passage score was used. The items based on each reading passage should be approximately unidimensional. However, still four DIF items were identified. The findings in the present study can be explained at least in part by changes in dimensionality of the criterion variable for measuring ability. One problem for using the passage is that there are only ten items in each reading passage. A short test may not produce reliable estimate of ability, which may influence the performance the MH and LR procedures. Rogers and Swaminathan (1993) indicate no test length effect on the MH and LR uniform indices. However, the shortest test they simulated was a 40-item test, which is the length of the total test used in this study.

The results of this study indicate little difference in the detection of uniform DIF between the MH and LR procedures. The agreement between the MH and LR procedures in the identification of uniform DIF is very high under all conditions. However, at smaller sample size, the difference between the two procedures is relatively large with the LR

procedure having a slight advantage. This is consistent with the results reported in Rogers and Swaminathan (1993), Swaminathan and Rogers (1990a, 1990b,) where the two procedures were almost equally effective in detecting uniform DIF. However, it should be noted that the frequency of DIF identification by the LR procedure is slightly higher than that by the MH procedure. Whether it indicates that the LR procedure is more powerful or that the false-positive rate of the LR procedure is higher needs further investigation. This finding is different from that in Swaminathan and Rogers (1990a, 1990b) where the MH procedure had a slight advantage in detecting uniform DIF. This is probably because the hypotheses of uniform and nonuniform DIF were tested simultaneously using a chi-square statistic with 2 degrees of freedom. When only uniform DIF was present, one degree of freedom was lost, hence the statistic might not be as effective as the MH procedure in detecting strictly uniform DIF. In the present study, the hypotheses of uniform and nonuniform DIF were tested separately using two chi-square statistics each with 1 degree of freedom. The slightly better performance of the LR procedure in the identification of uniform DIF is probably due to its greater power for this purpose, gained by conserving one degree of freedom.

The results of DIF detection by the LR nonuniform statistic indicated no items in this dataset exhibited serious nonuniform DIF in this dataset. Items 8, 10, and 29 produced relatively higher frequency of nonuniform DIF detection than other items but not high enough to deserve further investigation.

CHAPTER VI

CONCLUSION

The performance of the MH and LR statistics was investigated through replications using real data. The findings indicated that a sample size between 500 and 1,000 is necessary in order to obtain reliable MH and LR statistics. The results also indicate that there is substantial agreement between the MH and LR procedures in the identification of uniform DIF. The results of this study support the use of both the MH and LR procedures as practical means of detecting DIF. The ease of calculation and cost-effectiveness make the two procedures very attractive. Both procedures are very effective in detecting uniform DIF with the LR procedure having a slight advantage. However, when the magnitude and direction of DIF are desired, the MH procedure should be used. When the interaction between the ability level and group membership is suspected, the LR procedure is preferable. The advantage of the LR procedure in identifying nonuniform DIF is not very obvious in this study but has been shown elsewhere (Ibrahim, 1992; Rogers & Swaminathan, 1990a, 1990b; Swaminathan & Rogers, 1990; Tian, et al., 1994). Finally, this study suggests the importance of assessing the dimensionality of the test under scrutiny before doing DIF study. Test developers using the MH and LR procedures to assess DIF should be cautious in interpreting the results when the multidimensionality of the test is suspected.

This study provided information on the impact of dimensionality on the performance of the MH and LR procedures with real data. A suggestion for further research is that the

part dimensionality plays in DIF analysis may be further investigated with a simulation study so that the dimensionality of the test can be specified.

REFERENCES

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29(1), 67-91.

Ackerman, T.A., & Evanness, J.A. (1992). An investigation of the relationship between reliability, power, and the type I error rate of the Mantel-Haenszel and simultaneous item bias detection procedures. Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CAM, April 21-23, 1992). (ERIC 344937).

American College Testing Program: ACT Assessment Program, Technical Manual (1988).

Brown, P.C. (1992). An empirical study of the consistency of differential item functioning detection. Unpublished M.A. thesis. University of Ottawa, Ottawa, Ont.

Camilli, G. & Smith, J.K. (1990). Comparison of the Mantel-Haenszel test with a randomized and a jackknife test for detecting biased items. Journal of Educational Statistics, 15(1), 53-67.

Clauser, B.E., Mazor, K.M., & Hambleton, R.K. (1991a). Examination of various influences on the Mantel-Haenszel statistic. Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991). (ERIC 331876).

Clauser, B.E., Mazor, K., & Hambleton, R.K. (1991b). Influence of the criterion variable on the identification of differentially functioning test items using the Mantel-Haenszel statistic. Applied Psychological Measurement, 15(4), 353-359. (ERIC 331878).

Donoghue, J.R., & Allen, N.L. (1993). "Thin" versus "thick" matching in the Mantel-Haenszel procedure for detecting DIF. Journal of Educational Statistics, 18(2), 131-154.

Hambleton, R.K., & Cook, L.L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14(2), 75-96.

Hambleton, R.K., & Rogers, H,J. (1988). Detecting biased test items: comparison of the IRT Area and Mantel-Haenszel methods. Paper presented at the annual meeting of AREA, New Orleans, 1988. (ERIC 300398).

Hambleton, R.K., Rogers, H.J., & Arrasmith, D. (1988). Identifying potentially biased test items: a comparison of the Mantel-Haenszel statistic and several item response theory methods. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. (ERIC 296135).

Hambleton, R.K., & Swaminathan, H. (1985). Item Response Theory.Boston, MA: Kluwer-Nijhoff.

Hambleton, R.K., Swaminathan, H., & Rogers, H. (1991). Fundamentals of Item Response Theory. Newbury Park, California: SAGE

Hills, J. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8, 5-11.

Holland, P.W., & Thayer, D.T. (1986). Differential item functioning and the Mantel-Haenszel procedure. (ETS Tech. Rep. No. 86-69). Princeton, N.J.: Educational Testing Service.

Ibrahim, A.K. (1992). Distribution and power of selected item bias indices: A Monte Carlo study. Unpublished Ph.D. thesis. University of Ottawa, Ottawa, Ontario.

Kulick, E., & Hu, P.G. (1989). Examining the relationship between differential item functioning and item difficulty. College Board Report No. 89-5. ETS RR No.89-78. (ERIC 311 076)

Mazor, K.M., Clauser, B.E., & Hambleton, R.K. (1991). The effect of sample size on the functioning of the Mantel-Haenszel statistic. Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April, 1991). (ERIC 331877).

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

Ochieng, C.M.O. (1992). Examination of the distribution of the logistic regression and Mantel-Haenszel statistics under different condition of the null hypothesis: A Monte Carlo study. Unpublished M.A. thesis. University of Ottawa, Ottawa, Ont.

Pang, X.L., & Boss, M.W. (1993). The effects of sample size, item difficulty, and item discrimination on logistic regression item bias indices. Paper presented at the Annual Meeting of American Educational Research Association. Atlanta, Georgia.

Rock, D., & Chan, K. (1988). Differential item functioning analysis of math performance of Hispanic, Asian and White NAEP respondents. National Assessment of Educational Progress, ETS, Princeton, New Jersey. (ERIC 300440).

Rogers, H.J., & Swaminathan, H. (1993). A Comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17(2), 105-115.

Ryan, K.E. (1991). The performance of the Mantel-Haenszel procedure across samples and matching criteria. Journal of Educational Measurement, 28(4), 325-337.

Sudweeks, R.R., & Tolman, R.R (1990). The use of empirical versus subjective procedures for identifying science test items which function differentially for females and males. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching. Atlanta, GA. (ERIC 320770).

Swaminathan, H., & Rogers, H.J. (1990a). A comparison of the logistic regression and Mantel_haenszel procedures for detecting differential item functioning. Paper presented at the annual meeting of AREA, Boston, MA.

Swaminathan, H., & Rogers, H.J. (1990b). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4), 361-370.

Tian, F., Pang, X.L., & Boss, M. (1994). The consistency of the Mantel-Haenszel and logistic regression DIF identification procedures across sample size, over replications, and under different criteria. Paper presented at the Annual Meeting of American Educational Research Association. New Orleans, LA.

Wright, D.J. (1986). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. Paper presented at the annual meeting of the NCME. San Francisco.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? Journal of Educational Measurement, 15(3), 185-197.

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

APPENDIX A

Table 9 The Mean and Standard Deviation of MH-Z for Each Item Over 30 Replications

		Total Test		Subtest		Passage	
Items	Sample Size	Mean	SD	Mean	SD	Mean	SD
1	1,000	**.32**	.11	**.28**	.09	.17	.11
Total**	500	**.31**	.16	**.28**	.14	.17	.11
Sub1**	250	**.33**	.24	**.27**	.26	.18	.27
P1	100	**.46**	.31	**.40**	.27	.26	.27
2	1,000	**.26**	.07	**.21**	.07	.07	.09
Total**	500	**.24**	.09	**.20**	.08	.07	.09
Sub1*	250	**.26**	.13	**.23**	.13	.11	.15
P1	100	**.26**	.21	**.21**	.23	.05	.26
3	1,000	.16	.08	.10	.08	-.04	.09
Total	500	.16	.10	.12	.10	-.04	.09
Sub1	250	.25	.25	.11	.14	-.01	.14
P1	100	.22	.25	.13	.25	-.02	.24
4	1,000	.10	.07	.06	.07	-.06	.08
Total	500	.08	.10	.04	.09	-.06	.08
Sub1	250	.04	.14	.01	.14	-.12	.13
P1	100	.08	.19	.06	.17	-.08	.18
5	1,000	**.43**	.08	**.41**	.06	**.30**	.09
Total**	500	**.45**	.11	**.43**	.09	**.30**	.09
Sub1**	250	**.44**	.14	**.42**	.15	**.30**	.16
P1**	100	**.44**	.23	**.43**	.23	**.29**	.26
6	1,000	.14	.08	.09	.06	.00	.09
Total	500	.14	.10	.10	.09	.00	.09
Sub1	250	.13	.14	.11	.15	.00	.16
P1	100	.15	.20	.07	.21	-.06	.20
7	1,000	.10	.06	.05	.06	-.12	.07
Total	500	.11	.09	.05	.09	-.12	.07
Sub1	250	.06	.13	.02	.14	-.14	.15
P1	100	.08	.17	.06	.21	-.12	.19
8	1,000	.09	.06	.04	.06	-.09	.07
Total	500	.08	.09	.03	.09	-.09	.07
Sub1	250	.08	.12	.04	.13	-.10	.14
P1	100	.18	.17	.13	.17	-.04	.18
9	1,000	.05	.06	.01	.06	-.12	.07
Total	500	.02	.09	-.02	.09	-.12	.07
Sub1	250	.03	.12	-.01	.10	-.12	.12
P1	100	.03	.19	-.01	.19	-.15	.20
10	1,000	.15	.06	.12	.05	.00	.06
Total	500	.17	.07	.13	.07	.00	.06
Sub1	250	.16	.11	.13	.11	.01	.12
P1	100	.17	.19	.11	.17	.02	.18

Table 9 (Continued)

		Total Test		Subtest		Passage	
Items	Sample Size	Mean	SD	Mean	SD	Mean	SD
11	1,000	.00	.10	.06	.09	.18	.09
Total	500	.00	.12	.04	.14	.17	.13
Sub2	250	.02	.19	.08	.21	.20	.20
P2	100	-.02	.32	.06	.29	.24	.29
12	1,000	-.14	.09	-.08	.09	.01	.10
Total	500	-.17	.10	-.10	.10	-.01	.12
Sub2	250	-.14	.14	-.08	.14	.02	.14
P1	100	-.16	.28	-.09	.27	.01	.33
13	1,000	.11	.11	.19	.08	**.32**	.09
Total	500	.12	.11	.17	.11	**.31**	.12
Sub2	250	.12	.13	.17	.16	**.30**	.15
P2**	100	.14	.33	.23	.29	**.37**	.26
14	1,000	.08	.07	.12	.07	**.23**	.06
Total	500	.09	.07	.12	.07	**.23**	.09
Sub2	250	.10	.12	.15	.12	**.26**	.12
P2*	100	.05	.22	.11	.23	**.26**	.22
15	1,000	-.11	.08	-.08	.06	.03	.07
Total	500	-.13	.09	-.11	.10	.02	.10
Sub2	250	-.11	.11	-.08	.11	.03	.10
P2	100	-.13	.22	-.10	.19	.02	.19
16	1,000	-.19	.06	-.15	.05	-.03	.05
Total	500	-.18	.07	-.13	.07	-.01	.08
Sub2	250	-.20	.12	-.15	.11	-.03	.13
P2	100	-.18	.20	-.14	.17	.02	.17
17	1,000	**-.24**	.05	**-.21**	.05	-.11	.05
Total*	500	**-.23**	.06	**-.22**	.06	-.13	.07
Sub2*	250	**-.25**	.13	**-.22**	.13	-.13	.15
P2	100	**-.26**	.15	**-.26**	.13	-.15	.13
18	1,000	**-.20**	.05	-.16	.05	-.04	.05
Total*	500	**-.19**	.10	-.15	.10	-.02	.10
Sub2	250	**-.18**	.14	-.15	.13	-.02	.14
P2	100	**-.22**	.18	-.19	.17	-.06	.17
19	1,000	**-.36**	.06	**-.32**	.05	**-.21**	.06
Total**	500	**-.35**	.10	**-.31**	.09	**-.19**	.09
Sub2**	250	**-.35**	.10	**-.32**	.11	**-.21**	.12
P2*	100	**-.34**	.20	**-.33**	.20	**-.20**	.22
20	1,000	**-.23**	.08	**-.21**	.08	-.10	.09
Total*	500	**-.26**	.10	**-.23**	.12	-.12	.13
Sub2*	250	**-.24**	.15	**-.22**	.14	-.13	.15
P2	100	**-.28**	.32	**-.29**	.36	-.14	.37

Table 9 (Continued)

		Total Test		Subtest		Passage	
Items	Sample Size	Mean	SD	Mean	SD	Mean	SD
21	1,000	-.04	.09	-.09	.09	.07	.09
Total	500	-.05	.10	-.11	.10	.05	.11
Sub1	250	-.04	.19	-.08	.10	.05	.19
P3	100	-.07	.29	-.16	.31	.05	.09
22	1,000	-.09	.06	-.13	.06	-.05	.06
Total	500	-.08	.08	-.10	.09	-.03	.08
Sub1	250	-.08	.12	-.10	.11	-.03	.11
P3	100	-.09	.22	-.11	.22	-.03	.22
23	1,000	-.18	.07	**-.26**	.07	-.13	.07
Total	500	-.16	.09	**-.22**	.09	-.09	.09
Sub1**	250	-.20	.10	**-.28**	.10	-.16	.11
P3	100	-.19	.19	**-.29**	.19	-.11	.16
24	1,000	.01	.07	-.05	.06	.11	.06
Total	500	.00	.08	-.06	.07	.10	.07
Sub1	250	.01	.11	-.06	.11	.10	.11
P3	100	-.03	.20	-.10	.19	.09	.19
25	1,000	-.11	.06	-.19	.06	-.06	.07
Total	500	-.16	.08	-.22	.08	-.08	.08
Sub1	250	-.06	.14	-.12	.15	.01	.14
P3	100	-.11	.14	-.18	.16	-.03	.14
26	1,000	-.14	.06	**-.20**	.06	-.09	.06
Total	500	-.13	.09	**-.19**	.09	-.08	.10
Sub1*	250	-.15	.12	**-.22**	.12	-.10	.12
P3	100	-.24	.20	**-.25**	.21	-.15	.21
27	1,000	-.13	.07	-.18	.07	-.07	.07
Total	500	-.14	.09	-.20	.09	-.10	.09
Sub1	250	-.10	.11	-.14	.10	-.05	.11
P3	100	-.11	.17	-.15	.18	-.03	.19
28	1,000	.03	.06	-.02	.06	.12	.07
Total	500	.00	.09	-.04	.09	.10	.10
Sub1	250	-.02	.09	-.07	.10	.06	.11
P3	100	-.03	.18	-.03	.20	.11	.21
29	1,000	.03	.05	.00	.05	.10	.05
Total	500	.05	.08	.01	.08	.11	.08
Sub1	250	.04	.11	.01	.11	.10	.11
P3	100	.04	.20	.00	.19	.11	.18
30	1,000	-.04	.06	-.09	.06	.01	.06
Total	500	-.02	.09	-.07	.09	.03	.10
Sub1	250	-.04	.17	-.09	.15	.02	.17
P3	100	-.06	.16	-.10	.17	.01	.16

Table 9 (Continued)

		Total Test		Subtest		Passage	
Items	Sample Size	Mean	SD	Mean	SD	Mean	SD
31	1,000	-.04	.05	.02	.06	-.06	.07
Total	500	-.01	.10	.06	.11	-.03	.11
Sub2	250	-.04	.12	.02	.12	-.07	.12
P4	100	.02	.22	.08	.25	-.01	.23
32	1,000	.01	.06	.09	.06	.01	.05
Total	500	.01	.07	.08	.07	-.01	.10
Sub2	250	.01	.11	.08	.13	-.02	.15
P4	100	.03	.19	.15	.18	.03	.21
33	1,000	.05	.05	.13	.05	.03	.05
Total	500	.05	.09	.12	.08	.01	.08
Sub2	250	.06	.13	.14	.12	.03	.12
P4	100	.09	.20	.20	.19	.07	.19
34	1,000	.10	.07	**.20**	.08	.06	.09
Total	500	.14	.09	**.23**	.10	.09	.10
Sub2*	250	.15	.13	**.22**	.14	.09	.16
P4	100	.19	.23	**.26**	.24	.13	.28
35	1,000	.16	.05	**.23**	.06	.14	.06
Total	500	.13	.09	**.19**	.09	.10	.09
Sub2*	250	.12	.12	**.18**	.12	.10	.13
P4	100	.06	.23	**.14**	.21	.06	.22
36	1,000	.17	.07	**.27**	.08	.12	.08
Total	500	.17	.08	**.27**	.09	.11	.09
Sub2**	250	.16	.11	**.25**	.13	.11	.14
P4	100	.15	.22	**.26**	.23	.08	.24
37	1,000	.16	.06	**.24**	.06	.11	.07
Total	500	.16	.11	**.21**	.10	.08	.10
Sub2*	250	.19	.12	**.26**	.13	.14	.11
P4	100	.18	.22	**.26**	.22	.09	.19
38	1,000	-.11	.08	-.08	.06	-.17	.06
Total	500	-.10	.09	-.06	.10	-.14	.11
Sub2	250	-.16	.11	-.12	.11	-.20	.12
P4	100	-.14	.21	-.10	.20	-.19	.21
39	1,000	-.03	.05	.01	.04	-.08	.05
Total	500	-.02	.07	.03	.07	-.06	.07
Sub2	250	.00	.13	.05	.13	-.03	.13
P4	100	.00	.22	.03	.21	-.08	.20
40	1,000	-.03	.05	.02	.06	-.12	.05
Total	500	-.02	.07	.03	.08	-.11	.08
Sub2	250	-.04	.13	.02	.13	-.11	.13
P4	100	-.04	.18	.02	.25	-.14	.26

Note: * Possible DIF, ** Definite DIF

APPENDIX B

Table 10 The Frequency of DIF Identifications over 30 Replications by the MH and LR Chi-Square Statistics at the .01 Significance Level

Items	Sample Size	Total Test			Subtest			Passage		
		MH	LR(U)	LR(N)	MH	LR(U)	LR(N)	MH	LR(U)	LR(N)
1	1,000	**25**	**26**	1	**16**	**17**	0	2	3	0
Total**	500	**13**	**15**	0	**6**	**7**	0	2	3	0
Sub1**	250	**5**	**5**	0	**4**	**5**	0	4	5	0
P1	100	**1**	**1**	1	**0**	**2**	0	0	1	0
2	1,000	**26**	**28**	1	**19**	**19**	3	2	2	1
Total**	500	**13**	**14**	2	**9**	**9**	1	2	2	1
Sub1*	250	**9**	**10**	1	**6**	**8**	1	1	2	0
P1	100	**1**	**3**	0	**1**	**2**	0	1	2	0
3	1,000	10	10	3	3	5	5	0	0	0
Total	500	4	4	1	0	1	0	0	0	0
Sub1	250	0	1	1	0	1	1	0	0	0
P1	100	0	0	1	0	0	0	0	0	0
4	1,000	6	6	0	0	1	0	1	1	0
Total	500	0	1	0	0	0	0	1	1	0
Sub1	250	0	1	1	0	0	0	1	2	0
P1	100	0	0	0	0	0	0	0	0	0
5	1,000	**29**	**30**	4	**30**	**30**	2	**24**	**24**	0
Total**	500	**28**	**29**	3	**29**	**29**	1	**24**	**24**	0
Sub1**	250	**22**	**25**	2	**21**	**22**	1	**9**	**10**	0
P1**	100	**6**	**16**	0	**11**	**14**	0	**1**	**2**	0
6	1,000	8	7	0	4	3	0	0	0	0
Total	500	4	5	1	2	2	0	0	0	0
Sub1	250	3	3	1	1	2	0	0	1	1
P1	100	2	1	0	1	1	1	0	0	0
7	1,000	6	7	2	2	2	2	3	4	2
Total	500	3	2	0	2	1	1	3	4	2
Sub1	250	0	0	0	0	0	0	4	4	0
P1	100	0	0	0	0	0	1	0	0	1
8	1,000	5	4	6	1	1	10	2	2	4
Total	500	3	3	2	0	1	3	2	2	4
Sub1	250	1	2	1	0	1	1	1	1	0
P1	100	0	1	0	0	1	0	0	0	0
9	1,000	1	1	1	0	0	2	1	3	1
Total	500	1	1	1	0	0	0	1	3	1
Sub1	250	1	0	1	0	0	0	1	1	0
P1	100	0	0	0	0	0	1	0	0	0
10	1,000	15	15	7	6	6	4	0	0	4
Total	500	11	9	4	4	4	4	0	0	4
Sub1	250	3	5	4	0	2	2	0	0	5
P1	100	2	2	1	1	1	1	0	1	1

Table 10 (Continued)

Items	Sample Size	Total Test			Subtest			Passage		
		MH	LR(U)	LR(N)	MH	LR(U)	LR(N)	MH	LR(U)	LR(N)
11	1,000	2	1	0	1	1	0	7	8	3
Total	500	0	0	0	0	0	1	1	2	0
Sub2	250	0	0	0	0	0	0	2	2	0
P2	100	0	0	0	0	0	0	0	0	0
12	1,000	5	6	1	2	1	6	0	0	1
Total	500	3	3	1	0	1	2	0	0	0
Sub2	250	0	0	0	0	0	1	0	0	3
P2	100	0	0	0	0	0	1	0	0	0
13	1,000	2	2	0	7	7	1	**23**	**23**	3
Total	500	0	0	0	1	2	0	**9**	**10**	1
Sub2	250	0	0	1	0	1	0	**1**	**2**	0
P2**	100	0	0	0	0	0	0	**0**	**2**	0
14	1,000	2	2	0	6	6	0	**26**	**26**	2
Total	500	1	1	1	2	3	1	**14**	**14**	1
Sub2	250	1	1	0	3	3	0	**9**	**9**	0
P2*	100	0	0	0	0	0	0	**2**	**2**	0
15	1,000	11	11	2	2	2	2	1	1	1
Total	500	7	8	0	4	5	2	0	0	1
Sub2	250	1	2	0	0	2	0	0	0	0
P2	100	0	0	0	0	1	0	0	0	1
16	1,000	24	24	2	16	16	3	0	0	0
Total	500	11	12	0	4	5	1	0	0	1
Sub2	250	4	4	0	0	3	0	0	0	1
P2	100	0	1	0	0	0	0	0	0	2
17	1,000	**30**	**30**	3	**28**	**29**	1	6	7	0
Total*	500	**21**	**22**	1	**15**	**19**	2	1	4	1
Sub2*	250	**11**	**12**	0	**7**	**10**	0	2	4	1
P2	100	**0**	**4**	0	**1**	**2**	0	0	0	0
18	1,000	**25**	**25**	2	17	19	2	0	1	0
Total*	500	**14**	**13**	0	10	11	0	0	1	0
Sub2	250	**6**	**7**	0	4	6	0	0	0	0
P2	100	**0**	**1**	0	0	0	1	0	0	1
19	1,000	**30**	**30**	1	**30**	**30**	3	**24**	**26**	1
Total**	500	**26**	**29**	1	**25**	**27**	2	**10**	**13**	2
Sub2**	250	**18**	**19**	0	**15**	**18**	0	**2**	**4**	0
P2*	100	**4**	**7**	1	**5**	**6**	0	**2**	**3**	0
20	1,000	**18**	**18**	1	**12**	**17**	2	4	7	0
Total*	500	**12**	**15**	0	**8**	**12**	1	1	2	0
Sub2*	250	**1**	**2**	1	**1**	**2**	2	0	1	2
P2	100	**1**	**3**	0	**2**	**3**	0	1	1	1

Table 10 (Continued)

Items	Sample Size	Total Test			Subtest			Passage		
		MH	LR(U)	LR(N)	MH	LR(U)	LR(N)	MH	LR(U)	LR(N)
21	1,000	1	1	0	3	3	0	0	0	0
Total	500	0	0	0	0	0	0	0	0	1
Sub1	250	0	1	1	1	1	1	0	0	1
P3	100	0	0	0	0	0	0	0	0	1
22	1,000	2	3	0	9	8	0	1	0	0
Total	500	0	0	1	1	2	0	0	0	0
Sub1	250	0	0	0	0	0	0	0	0	0
P3	100	0	0	0	0	1	0	0	0	0
23	1,000	19	19	0	**28**	**29**	0	8	8	0
Total	500	5	5	1	**11**	**16**	1	1	1	1
Sub1**	250	3	6	0	**9**	**10**	0	3	3	1
P3	100	0	0	0	**2**	**3**	0	0	1	0
24	1,000	2	1	0	0	0	0	4	4	0
Total	500	0	0	0	1	1	0	0	1	0
Sub1	250	0	0	0	0	1	0	0	0	0
P3	100	0	0	1	0	0	2	0	0	0
25	1,000	10	11	1	20	21	0	2	2	1
Total	500	7	8	0	13	15	1	0	0	1
Sub1	250	1	2	0	2	2	0	1	1	0
P3	100	0	0	0	0	0	0	0	0	0
26	1,000	12	12	1	**19**	**22**	2	2	5	0
Total	500	4	7	1	**10**	**11**	0	2	4	0
Sub1*	250	2	2	0	**5**	**5**	1	1	1	2
P3	100	0	1	0	**0**	1	0	0	0	0
27	1,000	11	12	0	18	19	3	5	5	0
Total	500	7	9	0	13	13	0	3	3	0
Sub1	250	1	2	0	1	2	0	0	0	0
P3	100	0	0	0	0	1	0	0	0	0
28	1,000	1	1	0	0	0	0	7	9	1
Total	500	0	0	0	1	1	0	1	1	0
Sub1	250	0	0	0	0	0	2	0	0	0
P3	100	0	0	0	0	0	1	0	2	1
29	1,000	1	1	1	0	1	3	2	4	1
Total	500	0	0	0	0	1	1	0	2	1
Sub1	250	1	1	2	0	0	1	2	2	0
P3	100	0	1	1	0	0	1	1	1	1
30	1,000	1	2	1	5	7	2	0	0	2
Total	500	1	1	0	1	2	0	0	0	0
Sub1	250	1	1	0	1	1	0	2	2	1
P3	100	0	1	1	0	1	1	0	0	1

Table 10 (Continued)

Items	Sample Size	Total Test			Subtest			Passage		
		MH	LR(U)	LR(N)	MH	LR(U)	LR(N)	MH	LR(U)	LR(N)
31	1,000	1	1	2	0	0	0	0	3	2
Total	500	1	1	2	3	4	1	0	0	0
Sub2	250	0	0	1	0	0	2	0	1	0
P4	100	0	0	0	0	0	2	0	0	0
32	1,000	0	0	0	3	3	2	0	0	0
Total	500	0	0	0	0	1	0	0	0	0
Sub2	250	0	0	0	0	0	3	0	0	1
P4	100	0	0	0	0	0	0	0	0	0
33	1,000	0	0	0	7	7	1	0	0	0
Total	500	0	1	0	2	2	0	0	0	0
Sub2	250	0	0	0	1	2	1	0	0	0
P4	100	0	0	0	0	0	1	0	0	1
34	1,000	6	7	1	**21**	**22**	2	4	4	0
Total	500	4	4	0	**13**	**13**	0	0	1	1
Sub2*	250	4	3	0	**4**	**4**	1	2	2	0
P4	100	1	1	1	**2**	**3**	0	2	2	2
35	1,000	14	13	1	**26**	**27**	1	10	9	0
Total	500	5	5	0	**11**	**12**	0	2	1	0
Sub2*	250	2	1	0	**4**	**4**	0	0	0	0
P4	100	1	1	0	**1**	**2**	0	0	0	0
36	1,000	16	14	0	**27**	**26**	1	6	7	0
Total	500	3	3	1	**19**	**19**	2	1	1	0
Sub2**	250	0	1	0	**4**	**7**	1	0	0	0
P4	100	1	1	0	**1**	**2**	1	1	1	0
37	1,000	12	12	1	**25**	**27**	3	5	4	0
Total	500	8	6	1	**9**	**10**	1	1	1	2
Sub2*	250	2	3	0	**4**	**6**	1	1	1	0
P4	100	1	2	1	**2**	**4**	0	0	0	1
38	1,000	8	8	2	5	6	1	14	16	0
Total	500	2	3	0	2	3	0	8	9	0
Sub2	250	2	3	0	1	1	0	4	3	0
P4	100	1	0	0	0	0	0	1	1	0
39	1,000	0	0	2	0	0	5	1	1	1
Total	500	0	0	2	0	0	1	0	1	0
Sub2	250	0	0	2	0	0	1	0	1	0
P4	100	1	1	0	1	1	0	0	2	0
40	1,000	0	0	0	0	0	1	1	4	0
Total	500	0	0	0	0	0	1	1	1	0
Sub2	250	0	0	0	0	0	0	0	0	0
P4	100	0	0	1	0	0	1	0	2	0

APPENDIX C

Table 11 The Frequency of DIF Identifications over 30 Replications by the MH and LR Chi-Square Statistics at .05 Significance Level

Items	Sample Size	Total Test			Subtest			Passage		
		MH	LR(U)	LR(N)	MH	LR(U)	LR(N)	MH	LR(U)	LR(N)
1	1,000	**28**	**29**	3	**25**	**25**	1	7	8	0
Total**	500	**18**	**22**	1	**15**	**18**	1	7	8	0
Sub1**	250	**9**	**13**	0	**5**	**7**	1	5	5	1
P1	100	**6**	**9**	6	**4**	**5**	2	2	3	1
2	1,000	**29**	**30**	6	**26**	**26**	5	6	7	1
Total**	500	**20**	**23**	4	**15**	**14**	2	6	7	1
Sub1*	250	**13**	**15**	1	**12**	**12**	1	6	6	1
P1	100	**6**	**8**	1	**5**	**6**	0	2	5	1
3	1,000	19	19	12	8	8	11	2	2	4
Total	500	9	11	4	4	5	3	2	2	4
Sub1	250	4	4	5	1	1	6	0	0	2
P1	100	4	7	2	0	2	1	0	1	3
4	1,000	11	12	3	5	7	2	2	2	2
Total	500	3	6	1	3	3	0	2	2	2
Sub1	250	2	2	1	2	2	2	4	5	0
P1	100	0	3	2	0	0	0	0	0	1
5	1,000	**30**	**30**	8	**30**	**30**	9	**27**	**27**	2
Total**	500	**29**	**30**	6	**30**	**30**	5	**27**	**27**	2
Sub1**	250	**29**	**30**	4	**25**	**27**	3	**13**	**16**	1
P1**	100	**17**	**20**	2	**16**	**18**	4	**8**	**11**	0
6	1,000	15	19	1	7	9	1	2	2	1
Total	500	9	9	2	6	6	0	2	2	1
Sub1	250	4	5	5	6	5	1	3	2	1
P1	100	3	3	0	2	3	2	0	1	3
7	1,000	10	11	5	6	6	10	10	11	6
Total	500	7	6	4	2	2	3	10	11	6
Sub1	250	2	3	1	2	1	1	5	5	6
P1	100	0	3	3	1	3	3	0	1	3
8	1,000	9	8	13	2	2	18	6	6	4
Total	500	5	5	3	3	3	4	6	6	4
Sub1	250	4	3	5	3	3	3	3	3	3
P1	100	3	3	1	1	1	1	1	2	1
9	1,000	3	5	4	1	1	4	6	7	2
Total	500	1	1	1	1	1	1	6	7	2
Sub1	250	1	2	2	1	1	4	2	2	1
P1	100	0	0	1	0	0	2	1	3	0
10	1,000	23	25	9	14	12	11	0	0	15
Total	500	18	17	9	12	11	10	0	0	15
Sub1	250	10	11	10	6	10	7	1	1	8
P1	100	4	2	1	2	2	3	1	2	4

Table 11 (Continued)

Items	Sample Size	Total Test			Subtest			Passage		
		MH	LR(U)	LR(N)	MH	LR(U)	LR(N)	MH	LR(U)	LR(N)
11	1,000	3	2	2	2	2	2	12	14	7
Total	500	1	0	1	3	2	3	6	7	3
Sub2	250	2	2	2	2	3	5	3	3	1
P2	100	0	1	2	1	1	2	1	3	0
12	1,000	10	12	7	6	6	10	2	2	4
Total	500	6	8	5	3	3	5	0	0	4
Sub2	250	2	2	4	0	1	5	0	0	3
P2	100	0	3	1	0	1	3	1	2	3
13	1,000	6	7	2	17	17	1	**27**	**28**	6
Total	500	2	2	1	3	6	1	**16**	**19**	4
Sub2	250	1	2	3	2	2	2	**5**	**8**	2
P2**	100	0	2	0	2	3	1	**2**	**4**	1
14	1,000	6	5	1	11	11	2	**29**	**28**	8
Total	500	4	3	2	6	6	1	**20**	**20**	2
Sub2	250	2	3	0	5	7	0	**14**	**15**	0
P2*	100	1	2	1	3	4	1	**5**	**4**	5
15	1,000	18	20	4	11	11	8	5	4	2
Total	500	10	13	6	9	9	5	3	3	1
Sub2	250	3	7	1	3	3	1	0	0	0
P2	100	1	4	0	1	2	3	1	1	3
16	1,000	27	29	4	23	23	5	2	2	1
Total	500	16	18	1	9	12	1	0	1	2
Sub2	250	11	13	4	7	10	2	1	1	1
P2	100	4	7	3	2	3	3	0	0	2
17	1,000	**30**	**30**	6	**30**	**30**	5	12	18	0
Total*	500	**29**	**28**	5	**24**	**26**	5	8	11	1
Sub2*	250	**24**	**18**	2	**17**	**18**	3	8	9	2
P2	100	**12**	**10**	0	**6**	**6**	1	1	4	0
18	1,000	**29**	**30**	3	22	25	5	1	2	0
Total*	500	**16**	**17**	3	13	14	3	2	2	2
Sub2	250	**12**	**12**	0	9	9	0	2	2	1
P2	100	**5**	**7**	1	2	6	2	0	0	1
19	1,000	**30**	**30**	4	**30**	**30**	6	**28**	**29**	2
Total**	500	**29**	**29**	3	**28**	**29**	5	**16**	**18**	2
Sub2**	250	**24**	**26**	2	**23**	**23**	1	**12**	**12**	2
P2*	100	**12**	**14**	4	**10**	**12**	1	**3**	**5**	3
20	1,000	**25**	**25**	3	**23**	**23**	2	7	10	3
Total*	500	**17**	**19**	2	**16**	**18**	2	6	9	1
Sub2*	250	**6**	**9**	4	**5**	**8**	6	1	2	7
P2	100	**3**	**6**	4	**5**	**5**	2	3	4	5

Table 11 (Continued)

Items	Sample Size	Total Test			Subtest			Passage		
		MH	LR(U)	LR(N)	MH	LR(U)	LR(N)	MH	LR(U)	LR(N)
21	1,000	3	2	0	5	5	0	1	1	0
Total	500	0	0	0	3	3	0	0	0	2
Sub1	250	2	2	1	2	1	1	0	1	3
P3	100	0	0	2	0	2	1	1	1	4
22	1,000	9	8	0	14	13	0	3	2	1
Total	500	2	3	2	5	7	3	0	1	3
Sub1	250	1	3	2	3	4	2	1	1	0
P3	100	2	5	1	3	6	2	1	2	3
23	1,000	24	25	0	**30**	**30**	1	16	17	1
Total	500	10	13	2	**23**	**24**	2	3	4	3
Sub1**	250	10	11	2	**15**	**19**	2	3	3	1
P3	100	0	4	0	**5**	**6**	2	1	2	0
24	1,000	2	1	0	3	4	1	11	11	0
Total	500	1	1	0	1	2	0	2	3	0
Sub1	250	0	1	0	1	1	4	3	3	1
P3	100	1	1	2	0	1	2	3	5	0
25	1,000	15	17	5	27	27	4	4	5	1
Total	500	13	12	0	22	23	2	5	4	2
Sub1	250	3	3	0	4	4	2	3	4	0
P3	100	0	1	1	0	2	2	0	0	1
26	1,000	16	17	0	**27**	**27**	8	10	10	1
Total	500	10	13	3	**16**	**18**	4	6	6	1
Sub1*	250	4	6	2	**11**	**13**	5	3	3	2
P3	100	6	8	8	**6**	**8**	0	2	2	1
27	1,000	13	16	4	22	24	5	7	9	4
Total	500	12	13	0	17	18	0	6	7	0
Sub1	250	2	2	0	4	4	0	1	1	0
P3	100	0	3	0	3	3	0	1	1	2
28	1,000	2	2	2	2	2	2	13	14	2
Total	500	1	2	1	2	2	1	6	7	1
Sub1	250	0	0	1	1	1	2	1	1	1
P3	100	0	0	2	0	0	4	3	3	2
29	1,000	2	2	16	2	1	6	12	11	6
Total	500	1	2	4	2	2	2	10	11	3
Sub1	250	2	2	3	1	2	3	3	3	2
P3	100	1	2	5	1	1	3	2	3	3
30	1,000	3	4	2	10	11	2	1	1	5
Total	500	1	2	2	4	6	3	2	2	3
Sub1	250	2	2	1	3	3	2	3	4	3
P3	100	1	1	2	2	2	3	0	1	4

Table 11 (Continued)

Items	Sample Size	Total Test			Subtest			Passage		
		MH	LR(U)	LR(N)	MH	LR(U)	LR(N)	MH	LR(U)	LR(N)
31	1,000	3	3	3	2	3	1	5	9	2
Total	500	1	1	3	5	5	4	0	2	1
Sub2	250	1	0	3	0	0	3	1	1	4
P4	100	1	3	2	2	2	3	2	2	4
32	1,000	2	3	2	7	7	2	0	0	1
Total	500	0	0	4	2	3	1	1	1	2
Sub2	250	0	1	2	1	2	4	0	1	2
P4	100	0	0	1	0	1	5	0	0	3
33	1,000	1	3	1	17	17	5	0	1	1
Total	500	2	2	0	7	10	2	0	1	1
Sub2	250	3	2	1	3	4	1	1	0	0
P4	100	2	1	3	4	4	2	1	0	1
34	1,000	11	12	6	**24**	**24**	7	5	4	2
Total	500	11	12	1	**23**	**24**	3	2	2	3
Sub2*	250	6	5	5	**12**	**12**	6	4	4	2
P4	100	5	6	2	**6**	**8**	3	2	2	3
35	1,000	23	23	3	**29**	**30**	5	15	15	0
Total	500	9	7	0	**17**	**18**	2	5	5	0
Sub2*	250	5	7	1	**7**	**9**	2	3	3	1
P4	100	2	2	1	**2**	**4**	1	1	2	1
36	1,000	22	21	3	**30**	**30**	5	11	11	1
Total	500	11	15	4	**22**	**23**	5	2	3	1
Sub2**	250	4	6	2	**12**	**13**	2	1	2	0
P4	100	1	2	2	**3**	**6**	2	1	2	0
37	1,000	21	20	2	**27**	**27**	3	11	11	4
Total	500	11	10	1	**18**	**17**	3	3	3	2
Sub2*	250	6	8	1	**12**	**15**	1	4	5	2
P4	100	3	5	4	**4**	**7**	4	2	2	4
38	1,000	12	17	6	8	8	5	20	23	1
Total	500	6	9	2	4	6	0	11	12	3
Sub2	250	5	8	2	3	3	1	8	13	2
P4	100	3	5	5	2	1	1	2	6	0
39	1,000	2	1	8	0	0	10	3	6	4
Total	500	0	0	6	0	0	4	1	3	2
Sub2	250	3	2	4	1	4	4	1	1	5
P4	100	1	2	0	2	2	0	2	2	0
40	1,000	0	2	2	1	1	4	11	14	1
Total	500	0	0	5	0	0	1	3	7	1
Sub2	250	0	1	1	0	1	0	1	2	0
P4	100	0	2	3	3	3	3	3	3	0