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 Influence of Sample Size, Effect Size, and Percentage of DIF Items on the Performance of the Mantel-Haenszel and Logistic Regression DIF Identification Procedures

Michael Kennedy
Faculty of Education

Thesis submitted to
the School of Graduate Studies and Research
in partial fulfillment of the requirements for the
Master of Arts degree in Education

University of Ottawa

1994

ISBN 0-612-00610-7

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ABSTRACT

The frequent use of standardized tests for admission, advancement, and accreditation has increased public awareness of measurement issues, in particular, test and item bias. The logistic regression (LR) and Mantel-Haenszel (MH) procedures are relatively new methods of detecting item bias or differential item functioning (DIF) in tests. In only a few studies has the performance of these two procedures been compared.

In the present study, sample size, effect size, and percentage of DIF items in the test were manipulated in order to compare detection rates of uniform DIF by the LR and MH procedures. Simulated data, with known amounts of DIF, were used to evaluate the effects of these variables on DIF detection rates.

Data sets with equal ability distributions were generated for the reference and focal groups with five levels of sample size (100/100, 200/200, 400/400. 600/600, and 800/800). A 66 item test was used with two levels of percentage of DIF (9% and 18%). Two levels of item difficulty (b) (-.5 and .5) were combined with three levels of item discrimination (a) (.3, .5, and .7). The amount of DIF, or effect size, was measured by the difference between the b values of the reference and focal group (b value difference) and by the difference between the proportion correct scores of the reference and focal group (p-value difference). One of the purposes of the study was to determine which measure of effect size would more accurately predict the DIF detection rates. Four levels of b value difference (.2, .4, .6, and .8) were used and 24 p-value differences were computed as a result of four levels of b value difference, three levels of a, and two levels of b. Regression

models were obtained predicting MH detection rates, LR detection rates, and Δ_{MH}. The percentage of correctly identified uniform DIF items over 100 replications was reported for the MH and LR procedure; false positive rates were also reported.

DIF detection rates were quite similar for the MH and LR procedures; however, the LR procedure marginally outperformed the MH procedure under all conditions. The LR procedure produced slightly more false positives than the MH procedure. Sample size and effect size were found to have a positive effect on detection rates. P-value difference was a more accurate measure of effect size than b value difference and, as such, explained more variance in MH and LR detection rates. P-value difference explained 95% of the variance in Δ_{MH}. It was difficult to generalize the effect of b values on DIF detection rates; however, when b was held constant, larger a values were associated with increased DIF detection rates. Moreover, it was the combination of a, b, and b value difference that produce the p-value difference, which seemed to be the determining factor in the detection rates. Higher DIF detection rates were observed for the test with 9% DIF items than for the test with 18% DIF items.

In detecting uniform DIF, the LR procedure had a slight advantage over the MH procedure at the cost of increased false positive rates. P-value difference was definitely a more accurate measure of the amount of DIF than b value difference.

ACKNOWLEDGEMENTS

I am extremely grateful to my supervisor, Dr. Marvin Boss, for his guidance, patience, and encouragement throughout the research and writing of this paper.

I thank Dr. Bruno Zumbo and Dr. Marc Gessaroli for sharing with me their enthusiasm and knowledge of statistics and data analysis.

The staff of St. Patrick School have encouraged me during these past four years, my thanks to them.

Finally, I would like to thank my family who have been very supportive and patient throughout my years at university.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Tests are widely used in modern society. Important decisions such as acceptance and advancement in education, career advancement, and the provision of special services are often based on the results of test scores. Test developers are aware that scores on a test can be affected by many other variables in addition to the variable of interest, the subject's ability in the tested area. Test developers try to reduce these extraneous variables by maintaining a constant test-taking environment for all examinees by following rigorous rules of test administration and by thoroughly reviewing test items with panels of experts. However, whenever a test is designed to measure a large population, subgroups of that population will inevitably be present. It is important to ensure that members of one subgroup do not have an advantage over members of another subgroup on the test as a whole or on any given item. Often such an advantage favours majority group members over minority group members. Hence, the issue of bias in testing has become an important, yet controversial, area in the measurement field.

Bias at the test level is referred to as test bias. In this case, the test is often used as a criterion for decision making; yet, the performance of one group is not predicted as accurately as the performance of another. As a result one group has an advantage over the other. Item bias results when one subgroup of a population has an advantage over another on a given item. Specifically, an item is termed as biased if the probability of answering the item correctly is greater for one group than another group for members of equal ability. An item is considered to be unbiased if all individuals having an equal

underlying intended-to-be-measured ability have an equal probability of answering the item correctly regardless of group membership (Ackerman, 1992). Holland and Thayer (1986) suggest that modern approaches to item bias focus on the fact that different groups of examinees may react differently to the same test question. By examining these differences, test developers may gain new information on the test item and on the experiences and backgrounds of the different groups of examinees (Holland & Thayer, 1986).

Evidence of item bias is gathered empirically by calculating the probability of success on the item for individuals in a particular group and comparing this success rate to the success rate of other equally capable individuals from another group. If the success rates are significantly (meaningfully) different then the item is considered biased. However, the term "bias" has many different connotations; specifically, it may be assumed that one group has an advantage over another. In fact, the evidence gathered empirically gives no indication of whether the observed differences are legitimate parts of the construct being measured. 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). Therefore, a more neutral and accurate term, differential item functioning (DIF), rather than item bias, is preferred in recent years to describe the empirical evidence obtained in the investigation of bias (Hambleton, Swaminathan, & Rogers, 1991). The term DIF is used in this study.

One of the basic principles in DIF studies is the notion of comparing only comparable members of the two groups of interest (Holland & Thayer, 1986).

Comparability means similarity in certain characteristics which strongly relate to performance on the studied item: measures of ability for which the item was designed, courses taken or other measures of relevant experience, and membership in other groups. In common practice, criteria on which examinees from the two groups are matched include the test scores as these are readily available and usually measure the same ability as the studied item (Holland & Thayer, 1986). Without matching criteria, a simple difference between the performance on the item for the two groups could be calculated, resulting in measure of impact rather than DIF. Impact confounds differences in examinee ability with item characteristics and is of little use in attempting to identify items that may truly disadvantage some groups of examinees (Holland & Thayer, 1986).

Mellenberg (1982) defined DIF as either uniform or nonuniform. Uniform DIF exists when there is no interaction between ability level and group membership. Therefore, the probability of answering the item correctly is greater for one group than the other uniformly over all levels of ability (Swaminathan & Rogers, 1990). In contrast, nonuniform DIF occurs when there is an interaction between ability level and group membership. Thus, the difference in the probability of a correct answer for the two groups is not the same at all ability levels (Swaminathan & Rogers, 1990).

Many different statistical procedures exist to identify DIF items. The most theoretically preferred procedures are item response theory (IRT) based. IRT procedures involve determining the parameter(s) of the item characteristic curve (ICC) which represents the probability of responding correctly as a function of increasing ability (Hills, 1989). Although IRT approaches were theoretically preferred by most researchers during

the 1980s, drawbacks to IRT continue to exist. IRT methods require relatively large sample sizes to produce stable estimates of parameters (Camilli & Smith, 1990). Complex computer programs mean that IRT methods are not cost effective for identifying DIF. Furthermore, most IRT procedures do not have an associated test of significance. As a result of these drawbacks, researchers began searching for non-IRT methods that could identify DIF items to the same degree as IRT methods.

Presently, two DIF identification methods have been identified which have aroused much interest of measurement experts. Holland and Thayer (1986) proposed the use of the Mantel-Haenszel (MH) procedure as an alternative to IRT methods in identifying DIF for reasons of simplicity, availability of a significance test, relatively small sample size requirement, and cost effectiveness. Swaminathan and Rogers (1990) advocate the use of the logistic regression (LR) procedure for identifying DIF. The LR procedure, based on the logistic regression model, takes into account the continuous nature of the ability scale and is of use in identifying both uniform and nonuniform DIF (Swaminathan & Rogers, 1990).

In the following chapter a review of the literature related to the MH and LR procedures is presented.

CHAPTER II

LITERATURE REVIEW

A review of the literature is presented beginning with a description of the Mantel-Haenszel and the logistic regression procedures for identifying DIF. A review of the research regarding the two procedures follows, concluding with a summary of the findings. The purpose and specific research questions of this study are also presented.

The Mantel-Haenszel (MH) Procedure

The MH procedure was first proposed by Mantel and Haenszel (1959) for the study of matched groups in the medical field. Holland and Thayer (1986) identified the MH procedure as a measure of item bias or differential item functioning. The MH procedure involves comparing examinees from two groups. Typically the group of interest (the focal group) is compared to the larger group (the reference group) on each item in the test. The item of interest is termed the studied item. Examinees from both groups are matched on a criterion strongly related to performance on the studied item. Usually, the total test score is used as this criterion, although other measures of the ability for which the item is designed may be used along with measures of relevant experience or courses taken in the area of interest. Once the matching criterion(ia) has(ve) been selected, data for the studied item for the examinees in the reference group and the focal group are arranged into a series of 2x2 contingency tables, one for each level of the matching criterion (see Table 1).

Table 1: Frequencies of responses of focal and reference groups at a given ability level. (Holland & Thayer, 1986)

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

In Table 1, T_j is the total number of examinees in both the reference and focal groups in the j^{th} matched score interval; n_{Rj} is the number of examinees in the reference group whereas n_{Fj} is the number of examinees in the focal group; A_j and C_j represent the number of examinees who gave a correct response to the studied item in the reference and focal groups, respectively; B_j and D_j represent the number of examinees who gave an incorrect response to the studied item in the reference and focal groups, respectively; m_{1j} and m_{0j} are the total number of examinees who responded to the item correctly and incorrectly, respectively.

The Mantel-Haenszel alpha (MH-alpha)

Using the contingency tables, the ratio of the odds for success of the reference and focal group members is calculated for each score interval, weighted according to the number of individuals in that interval. The MH statistic is calculated by taking an average of these odds ratios across the criterion (score scale). This common odds-ratio of success, known as MH-alpha, can be defined as follows:

$$\alpha_{MH} = \frac{\Sigma\, A_j D_j / T_j}{\Sigma\, B_j C_j / T_j}$$

MH-alpha can also be defined using probability notation:

$$\alpha_{MH} = \frac{p_{Rj}}{q_{Rj}} / \frac{p_{Fj}}{q_{Fj}} = \frac{p_{Rj} q_{Fj}}{p_{Fj} q_{Rj}} \qquad \text{for all } j = 1,\ldots,k$$

where p_{Rj} and q_{Rj} are the probabilities of answering the item correctly and incorrectly, respectively, for reference group examinees in the j^{th} interval; similarly, p_{Fj} and q_{Fj} are the corresponding probabilities for focal group examinees.

This ratio has a scale of 0 to infinity. When $\alpha_{MH} = 1$, the odds for success on an item are the same for the reference and focal groups; therefore, DIF is not identified. When the odds for success on an item differ substantially between the reference and focal groups, DIF is identified. When α_{MH} is greater than 1, the likelihood that reference group members got the item correct exceeds the corresponding likelihood for comparable focal group members. When α_{MH} is less than one, the item favours the focal group.

The Mantel-Haenszel chi-square (MH-CHISQ)

A chi-square test of significance is associated with α_{MH}. Using the same probabilities defined above, the MH chi-square tests the following null hypothesis:

$$H_0 : \frac{p_{Rj} / q_{Rj}}{p_{Fj} / q_{Fj}} = 1 \qquad \textit{for all } j=1,\ldots,k$$

versus

$$H_1 : \frac{p_{Rj} / q_{Rj}}{p_{Fj} / q_{Fj}} \neq 1 \qquad \textit{for all } j=1,\ldots,k$$

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

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

where

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

and

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

A significant MH chi-square statistic suggests that the item functions differently for

the two groups of examinees. Under the H_0 the MH chi-square has an approximate chi-square with one degree of freedom. The alternative hypothesis is nondirectional; consequently, the MH chi-square can identify DIF that favours either subgroup.

The Mantel-Haenszel Delta (Δ_{MH}) and Mantel-Haenszel-Z (MH-Z)

Holland and Thayer (1986) proposed taking a log transformation of α_{MH} to put it onto a symmetric scale in which 0 is the null value. This log transformation, known as Δ_{MH}, was suggested by Holland and Thayer (1986) to put α_{MH} on a scale similar to the scale of differences in item difficulty used by the Educational Testing Service (ETS). The ETS delta scale ($\bar{x}$=13, SD=4) results from a linear transformation of z which is estimated from a classical definition of item difficulty, p. A log transformation is used to put α_{MH} on the ETS delta scale with zero as the null value:

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

The value of Δ_{MH} is the average amount more difficult an examinee in the reference group found the item than did comparable examinees in the focal group. Negative values of Δ_{MH} correspond to items that the reference group found easier on average than did comparable focal group members. When Δ_{MH} is positive, the item is easier for focal group members. For medium difficulty items, a Δ_{MH} of 1 indicates a difference in item difficulty, p, of about .10. The Δ_{MH} provides a measure of the magnitude and direction of DIF in studied items in terms of the delta scale of item difficulty employed by ETS.

A z-score distribution has a mean of zero and standard deviation of one. Since Δ_{MH} has a standard deviation of 4, the Δ_{MH} scale and a z-score scale have standard deviations in the ratio of four to one. Δ_{MH} can therefore be expressed in terms of a scale of z-scores by simply dividing by four, which produces the MH-Z:

$$MH\text{-}Z = -(1/1.7)\ \ln(\alpha_{MH})$$

Both Δ_{MH} and MH-Z have a value of 0 under the null hypothesis and give a measure of the direction and magnitude of DIF in the studied item. Δ_{MH} provides a measure in terms of the ETS difficulty scale whereas, the MH-Z does so in terms of a z-scale. Thus, for medium difficulty items, an MH-Z of .25 indicates a difference in item difficulty, p, of about .10. If MH-Z<0, the reference group performed better than the focal group. If MH-Z>0, the focal group performed better than the reference group.

The Logistic Regression (LR) Procedure

Based on the logistic regression model, the LR procedure takes into account the continuous nature of the ability scale and can be used to identify both uniform and nonuniform DIF (Swaminathan & Rogers, 1990). The probability of a correct response to an item from given independent variables is predicted by the logistic regression model as follows:

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

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.

By specifying separate equations for the two groups of interest (reference and focal) the logistic regression model given above can be used to model differential item functioning:

$$P(u_{ij} = 1 \mid \theta_{ij}) = \frac{e^{(\beta_{0j}+\beta_{1j}\theta_{ij})}}{[1+e^{(\beta_{0j}+\beta_{1j}\theta_{ij})}]}, \quad i=1,\ldots, n_j, \quad j=1,2.$$

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

No DIF is present if the logistic curves for the two groups are the same; that is, if $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$. When $\beta_{11} = \beta_{12}$ but $\beta_{01} \neq \beta_{02}$, the curves are parallel but not coincident and hence uniform DIF is present. When $\beta_{01} = \beta_{02}$ but $\beta_{11} \neq \beta_{12}$, the curves are not parallel and hence nonuniform DIF is present (Swaminathan & Rogers, 1990). Nonuniform DIF is also present if $\beta_{01} \neq \beta_{02}$ and $\beta_{11} \neq \beta_{12}$.

As with the MH procedure, the LR procedure also uses a chi-square test of significance to test the hypothesis of no DIF:

$$H_0 = \beta_{01} = \beta_{02} \text{ and } \beta_{11} = \beta_{12}$$

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. This results in two chi-square statistics each with one degree of freedom, for uniform and nonuniform DIF respectively.

Studies of the MH and LR Procedures

Many researchers in the area of DIF have investigated the MH procedure for identifying biased items. In recent studies, researchers have compared the MH procedure to other DIF identification procedures, namely IRT methods. As well, researchers have manipulated variables to determine the effectiveness of the MH procedure. Some variables of interest are sample size, test length, item discrimination, item difficulty, and choice of criteria for matching subjects. More recently the LR procedure has emerged in the literature as another method for identifying DIF. Many researchers are still studying the implications of the LR procedure, particularly, its agreement with the MH procedure. In the following studies, the MH procedure is examined. In some studies, variables are manipulated in order to compare the rates of DIF detection between the MH and LR procedures.

Mazor, Clauser, and Hambleton (1991) used simulated data to investigate the effect of sample size on the performance of the MH procedure. A three parameter logistic model was used. Three data sets of 2000 examinees each were generated to allow for

comparisons of groups with equal ability, and of groups where the focal group was less able. Examinee ability in all three data sets was normally distributed with a standard deviation of 1. The first two data sets represented the Reference Group 1 and Focal Group 1 and had distributions with a mean of 0 to allow for comparisons between groups with equal ability distributions. To allow for comparisons between groups with unequal ability distributions, the mean for the third distribution, Focal Grcup 2, was set at -1.0.

Five different 75 item tests were generated each having 59 common non-DIF items. Eighty DIF items were generated (five groups of 16 items) to combine with the 59 non-DIF items to comprise each 75 item test. Each group of 16 DIF items had four levels of item discrimination (a): .25, .60. .90, and 1.25 crossed with four differences in b between the reference and focal group: .25, .50, 1.00, and 1.50. These differences in b simulated various degrees of DIF. Finally, five values of b were used for the reference group (-2.5, -1.0, 0, 1.0, and 2.5); one for each group of 16 items. Within the entire set of 80 DIF items, each level of b was completely crossed with each level of a and with each difference in b-value. The c value was held constant at .20.

The MH coefficient was computed for sample sizes of 2000, 1000, 500, 200, and 100 examinees in both the reference and focal groups. In order to minimize the impact of chance variability, the 500 run was replicated once for each set and the 200 and 100 runs were replicated twice. Results were reported at the .01 level of significance.

The percentage of DIF items correctly identified decreased as the sample size decreased. With sample size of 2000, 64% and 74% of the DIF items were correctly identified for the unequal and equal ability distributions, respectively. However, with

sample size of 100, this decreased to 9% and 18% respectively. DIF identification rates were consistently higher with equal ability distributions as compared to unequal ability distributions.

Items with low a values were rarely identified, requiring larger sample sizes and greater differences between the two groups on item difficulty. Very difficult items were also rarely flagged by the MH procedure since very few examinees actually got these difficult items correct. Items with larger b differences (more DIF) were more likely identified than those with smaller b differences.

Of practical concern is the amount of DIF missed with small sample sizes. The authors believe that the results of the MH procedure are questionable at small sample sizes Only where the most markedly DIF items are a concern would sample sizes of 200 be considered adequate. The authors recommend using sample sizes of at least 1000.

In a similar study, Clauser, Mazor, & Hambleton (1991a) examined the effects of changes in item difficulty (b), item discrimination (a), and amount of simulated DIF on the performance of the MH procedure. This study differed from the one above (Mazor, Clauser, & Hambleton, 1991) in that sample size was not changed, but rather, set at 1000 for both groups. In particular, the authors were interested in determining the statistical characteristics of items with known DIF that were undetected by the MH procedure. Five 75-item tests (16 DIF items, 59 common non-DIF items) were used varying item difficulty and discrimination for the 16 DIF items as above (Mazor, Clauser, Hambleton, 1991). Equality of ability distributions was also studied as above. Thus, there were potentially 80 DIF items that could have been correctly identified by the MH procedure.

At the .01 level of significance the MH procedure identified 49 of the 80 items that had some level of simulated DIF. Results showed that as the difference in the difficulty for focal and reference groups increased, the probability that the item would be identified as DIF increased dramatically. Increases in item discrimination also increased the likelihood of DIF identification. A relationship between item difficulty and discrimination was most apparent. Low discrimination was associated with low MH chi-square values even when the difference in item difficulty was large. Highly discriminating items produced high MH chi-square values when compared across various item difficulty values. Results also indicated that the MH procedure did not detect DIF in the most difficult items as there were too few examinees at the upper end of the ability distributions. Thus, the effect of high item difficulty was limited to the extreme upper range of the ability scale. Furthermore, the MH procedure was more sensitive with groups of equal ability distributions. The continued use of the MH procedure for DIF identification was supported.

Clauser, Mazor, & Hambleton (1991b) investigated the effectiveness of the MH procedure in detecting DIF test items while varying the matching criterion. The results of the MH statistic were compared across different criteria. The authors predicted that by systematically grouping the test items into different subtests and then using these various subtest scores as the matching criterion, the effectiveness of the MH procedure in identifying DIF would be influenced.

Data for the study came from the 1982 administration of the New Mexico High School Proficiency Exam, with responses of 23,000 students to 150 test items. Two groups were selected for comparison: 8,000 Anglo-American and 2,600 Native American

students. A random sample of 1,000 examinees each was selected from both Anglo-American and Native American subjects.

For the combined sample of 2,000 examinees, items with very low discrimination and low difficulty were removed. The authors argued that there was little merit in analyzing test items for DIF if the items were contributing very little to the test score variability. This left a total of 91 items which were divided into three subtests. Items were randomly assigned to tests with the stipulation that each test contain 75 items and each item be represented in at least two of the three tests. The MH procedure was then carried out for each of these three tests, separately, using the 75 item test score as the matching criterion. Items were identified as DIF only if they were identified as DIF in all of the tests in which they appeared.

Next the items were categorized as belonging to one or more subtest based on skills required to answer the items correctly: Math (27 items), Reading (15 items), Prior Knowledge (49 items), and Charts (19 items). The first three skill subtests were mutually exclusive, whereas items in the Charts subtest were coded in one of the other skill subtests as well. Using the skill subtest score as the matching criterion, the items were tested for DIF.

Finally, three additional subtests were constructed by randomly assigning each of the 91 items to one of three subtests with test lengths approximately equal to the skills subtests formed. This resulted in three control subtests of 30, 31, and 30 items. The items were again tested for DIF using the control subtest score as the matching criterion. The

authors could now evaluate the extent to which a reduction in the number of items analyzed would affect the MH procedure.

Based on three runs of the MH computer program analyzing the 91 items (randomly assigned to three tests of 75 items each), 22 items met the conditions of DIF between the Anglo-American and Native American groups. When the four skill subtests (Math, Reading, Prior Knowledge, and Charts) were analyzed, a number of changes in the MH results were observed. One third of the DIF test items (7 of 22) ceased to be DIF when analyzed within the skills subtests, whereas 10 previously non-DIF items were now identified as DIF. The authors' hypothesis was supported in that changes in item grouping affected the results.

The three randomly selected control subtests of 30, 31, and 30 items, which provided subtests similar in numbers of items to the skill subtests but without the constraints on the skills measured, were analyzed for DIF. All of the 22 items previously identified as DIF with the larger test, continued to be identified in the control subtests. A less easily predicted phenomenon was also noted as a result of the subtest runs: 12 items previously not identified as DIF with the larger test were identified in the control subtests. This was not expected.

In summary, the authors suggest that test developers using the MH procedure to assess item bias in tests should be cautious in interpreting the results. It appears that the context in which items are studied can influence the results. As well, the results show that as the length of the test decreased, the number of additional items identified as DIF

increased. Further research is needed to determine the extent to which these results can be generalized to other test data and to provide a theoretical explanation of the results.

Miller and Oshima (1992) used a two-stage procedure for estimating DIF with the MH procedure and with six IRT indices including the signed area (SA) and unsigned area (UA), the signed and unsigned sums of squares (SSOS and USOS), and the weighted signed and weighted unsigned sums of squares (WSSOS and WUSOS). The authors varied the sample size, the number of DIF items, and the magnitude of DIF. The two levels of sample size for the reference and focal groups respectively were 1000/1000 and 1000/300. Forty item tests were simulated with four levels of the percentage of DIF items: 5%, 10%, 20%, and 40%. The magnitude of DIF was measured by the difference in the b values between the reference and focal groups: small DIF (b difference of .20), moderate DIF (b difference of .35), and mixed DIF (b difference of .20 and .35). Baseline estimates of all the statistics were used to decide whether an item had DIF or not. An item was identified as DIF if it exceeded the mean plus two SDs from the distribution of indexes in the baseline comparison. In Stage 1 of the procedure, items were compared to the baseline conditions. In Stage 2, the DIF items from Stage 1 were removed and the baseline procedure was followed on the reduced set of items. Essentially the two-stage procedure involved identifying DIF, then removing the DIF items and reestimating the DIF.

Of interest here is the performance of the MH procedure. The MH procedure identified moderate DIF as well as the IRT indices did, and the MH statistic identified fewer false positives across all conditions. However, the MH procedure was not as powerful as the IRT indices in identifying items with small DIF. As expected, the IRT procedures and

the MH procedure identified more moderate DIF items than small DIF items. For the equal sample size run (1000/1000), as the percentage of moderate DIF items increased, the number of DIF items identified by the MH procedure increased proportionately. Since Stage 2 identified very few additional DIF items when the number of DIF items was small and the magnitude of DIF was weak, the authors concluded that the two-stage procedure for DIF identification may be less useful than single-stage procedures when identifying ethnic, race, or gender bias.

Donoghue and Allen (1993) examined the effects of the matching variable on the identification of DIF with the MH statistic. Thin matching implied matching subjects on each value of the criterion variable, for example, matching on total test score. Thick matching involved restricting the number of raw score categories to create fewer intervals.

Six independent variables were manipulated in the study: amount of DIF (no DIF, DIF favouring the reference group), method of forming the matching variable (11 levels), test length (5, 10, 20, and 40 items), sample size RG/FG (300/100, 600/200, and 1200/400), item discrimination (0.3, 1.0, and 1.5), and item difficulty (-1.5, -1.0, -0.5, 0.0, 0.5, 1.0, and 1.5). Each test was replicated 20 times. Since the study was exploratory in nature many observations (up to 27,720) were made as each of the six independent variables was crossed with each other in order to determine whether thick matching might provide an advantage over thin matching.

In general, the results were as follows. Thick matching improved the performance of the MH procedure for short tests (5 to 10 items) whereas thin matching tended to increase the rate of false positives. However, for long tests (40 items), especially, with

adequate sample sizes (1600/400), thin matching yielded the best results compared to any thick matching method examined.

In developing the thick matching methods, Donoghue and Allen (1993) created intervals of the total score. The authors suggest that perhaps using auxiliary information in addition to the total score may be seen as a finer degree of matching than matching on total score alone. Matching the examinees on other information as well as test score may improve the degree of matching and should help to eliminate some items identified as DIF.

Indeed, the use of auxiliary information in determining the matching variable was studied by Zwick and Ercikan (1989). The authors hypothesized that the number of DIF items would decrease when auxiliary information was used in addition to the test score to match subjects.

Data came from the 1986 National Assessment of Educational Progress (NAEP) survey. In particular, DIF analyses were based on the U.S. History Assessment administered to a national sample of 7,743 Grade 11 students. There were four U.S. History blocks (subtests) consisting of 34 to 36 cognitive items each. Within each of the four history subtests, DIF analyses were conducted to compare the performance of males and females, whites and blacks, and whites and Hispanics using the subtest score as the criterion. Sample sizes of the five groups varied slightly among the four History blocks, H1 to H4, as follows: males (964, 945, 935, and 1018), females (948, 984, 975, and 933), whites (1375, 1365, 1346, and 1410), blacks (321, 330, 306, and 308), and Hispanics (198, 168, 201, and 185).

The history items were classified into three categories (A, B, or C) according to the ETS rules which use the Mantel-Haenszel delta difference (Δ_{MH}) to identify DIF. 'A' items were considered to be free of DIF (Δ_{MH} absolute value less than 1). 'B' items have some DIF but could still be used as test items (Δ_{MH} absolute value between 1 and 1.5). 'C' items should be avoided as test items (Δ_{MH} absolute value greater than 1.5).

The following results occurred when the subtest score alone was used as the matching variable. For the male-female analysis, five items were classified as 'C' items: four items favoured males and one item favoured females. Twenty-five items were classified as 'B' items: 12 items favoured males and 13 items favoured females. In the white-black analysis there were three 'C' items, each favouring blacks. Of the 22 'B' items, 15 favoured whites and seven favoured blacks. In the white-Hispanic analysis one 'C' item was present favouring Hispanics. Ten 'B' items favoured Hispanics while 15 'B' items favoured whites.

After the initial identification of DIF items, subjects were classified according to the number of historical periods they claimed to have studied (Periods of Study). The number of Periods of Study had a strong relationship to overall performance on the NAEP history assessment. By matching of Periods of Study as well as score, examinees should be more closely matched. As such, the authors expected that this refinement in matching would produce fewer items showing DIF.

Results of the second analysis were nearly identical to the first analysis. As such, the more refined matching procedure produced at least as many DIF items as did matching

on score alone. The authors concluded by suggesting reasons why these unexpected results occurred.

Beginning with Swaminathan and Rogers (1990), in the next group of studies the performance of the MH and LR procedures is compared. Swaminathan and Rogers (1990) presented a logistic regression model for characterizing differential item functioning between two groups as outlined earlier. Included in their study was a comparison of the LR and MH procedures across sample size, test length, and the nature of DIF.

The authors generated data for six conditions, resulting from two levels of sample size (250 per group and 500 per group) with three levels of test length (40 items, 60 items, and 80 items). Within each test 20% of the items showed DIF, half with uniform DIF and half with nonuniform DIF. The nonuniform DIF simulated represented disordinal interaction between ability and group membership. The LR and MH procedures were compared with respect to the percentage of items that was correctly identified and the percentage of false positives.

For the items with uniform DIF, the two procedures had similar detection rates; the MH procedure had a slight advantage. Uniform DIF was detected with 75% accuracy by both procedures in samples of 250; while in samples of 500, DIF was detected with 100% accuracy by both LR and MH procedures. Furthermore, the LR procedure detected nonuniform DIF with 50% accuracy in small samples with short tests and with 75% accuracy in large samples with long tests. The MH procedure was unable to detect nonuniform DIF under any condition. The MH procedure performed better than the LR

procedure in terms of false positives producing 1% false positives in all conditions compared with 1% to 6% false positives for the LR procedure.

In summary, the authors show that the LR procedure is as powerful as the MH procedure in detecting uniform DIF and more powerful in detecting nonuniform DIF of a disordinal nature. According to Swaminathan and Rogers (1990) the main advantage of the LR procedure is that it provides a model-based approach for studying DIF and therefore may be more helpful in determining the nature of DIF than the MH procedure.

In further research, Rogers and Swaminathan (1993) implemented two simulation studies to observe the effectiveness of the LR and MH procedures for detecting DIF. The distributions of the test statistics were examined in the first study; in the second study, the relative power of the two procedures to detect uniform and nonuniform DIF was investigated.

In their investigation of the distributions of the LR and MH statistics, two factors were manipulated, sample size and the degree of model-data fit. Two levels of model-data fit ("good" fit and "poor" fit) were crossed with two levels of sample size (250/250 and 500/500). Good fit data were generated using a two-parameter IRT model; whereas, a three-parameter IRT model was used to generate poor fit data. Response patterns were generated from a 40-item test with item parameters selected to produce an approximately normal distribution of test scores. The same item parameters were used for each group; hence, all items were unbiased. For each combination of model-data fit and sample size, 100 replications were performed. Of the 40 items, five were chosen to vary in level of difficulty and discrimination. For each of these five items, the LR and MH test statistics

were calculated, and empirical sampling distributions were constructed. The Kolmogorov-Smirnov test was used to determine if the test statistics had the expected distributions.

Overall, the distributional assumptions of both procedures were met to a satisfactory degree. The one exception that occurred with the LR procedure was with an item of high difficulty and high discrimination. The authors concluded that the distributional assumptions of both procedures were met.

In the power study Rogers and Swaminathan (1993) manipulated factors that were assumed to affect the power of the LR and MH procedures. Thirty-two conditions were simulated by crossing two levels of model-data fit ("good" fit and "poor" fit), two levels of sample size (250/250 and 500/500), 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 (15% including the item of interest and 0% other than the item of interest). Within each condition, Both uniform and nonuniform DIF were simulated. Four sizes of DIF were studied with area values of .2, .4, .6, and .8 for each type of DIF in each condition.

Results supported the hypothesis that both the LR and MH procedures were almost equally effective in detecting uniform DIF. The LR procedure was more effective than the MH procedure in identifying nonuniform DIF. The MH procedure was sensitive only to DIF that is approximately constant across all trait levels; whereas, the LR model was sensitive to DIF of a more general nature. Larger samples increased the DIF detection rates, as did larger sizes of DIF. For both procedures, test length and shape of the score distribution did not appear to affect the detection rates.

Brown (1992) examined the issue of sample size in DIF detection over replications using both the MH and LR procedures. Data for the study came from the 1988 Reading and Written Expression assessments used to test approximately 100,000 British Columbia students in grades 4, 7, and 10. The Form M Achievement Survey item responses were used. Twenty test items assessed literal comprehension and 16 test items assessed inferential comprehension.

Responses of 33,809 grade four public school students formed the dataset. The student population was examined on the basis of two cultural factors, gender and regional location. Various comparison groups were used in this study to create a variety of conditions under which the detection rates and effects of sample size for the MH and LR indices could be studied. All comparison groups were created by randomly selecting cases by gender and/or region across six levels of sample size: 1000/1000, 750/750, 500/500, 300/300, 200/200, and 100/100 examinees per group.

Replications were carried out at each sample size where sufficient numbers were available. Items were flagged as statistically significant ($p<.05$) by the MH chi-square, LR uniform, and LR nonuniform chi-square procedures. These items were then reviewed and classified as DIF items according to prespecified standards.

Overall, the DIF detection rates were low for the MH and LR indices. Agreement between the MH chi-square and the LR uniform chi-square statistics in identifying items with DIF was high. The detection rates of both the MH and LR procedures were affected by sample size. Larger sample sizes had larger DIF detection rates.

Using real data from the ACT Assessment, Pang, Tian, & Boss (1994) examined the consistency of the MH and LR statistics across different sample sizes using either total test score or subtest score as the criterion over replications. The study was carried out on a population of Caucasians (107,502 females and 75,854 males) who wrote the ACT Assessment (Form 39B) in 1989. Item responses to the Mathematics test formed the data used in the study. The Mathematics test includes three subtests: Elementary Algebra (24 items, Subt1), Intermediate Algebra/Coordinate Geometry (18 items, Subt2), and Plane Geometry/Trigonometry (18 items, Subt3). Four sample sizes (1000/1000, 500/500, 250/250, and 100/100) were used for each procedure. Thirty replications were carried out under each combination of sample size and criterion.

The authors classified DIF into three levels: Definite DIF, Probable DIF, and Possible DIF. Based on the sample size of 1000, items with a mean value of MH-Z equal to or greater than an absolute value of .25 were classified as Definite DIF; items with an absolute mean value of MH-Z between .20 and .25 were classified as Probable DIF; and items with an absolute mean value of MH-Z as large as .15 but less than .20 were classified as Possible DIF.

Using the total test score as the criterion, thirteen items were identified as uniform DIF and two as nonuniform DIF: four Definite DIF, four Probable DIF, and seven Possible DIF. For the fifteen DIF items identified, the identification rates for each procedure increased markedly as the sample size increased for both total test score and subtest scores, indicating a strong effect of sample size. The authors suggest that to detect

Definite DIF a minimum sample size of 500 is necessary. For Probable and Possible DIF, a sample size of 1000 would be desirable.

Furthermore, the MH and LR procedures showed very similar detection rates for uniform DIF. LR proved to be slightly more powerful than MH in detecting uniform DIF; however, this effect was greater for small sample sizes.

When the criterion variable was examined, the performance of the two procedures was affected. In general, the detection rates and MH-Z decreased when the subtest score was used as the criterion. For two nonuniform DIF items, using the subtest score as the criterion drastically decreased the rate of DIF identification.

In summary, the authors concluded that the MH and LR procedures are virtually exchangeable in identifying uniform DIF. Nonuniform DIF was identified much less than uniform DIF. Because of the short subtest length, the effects of choosing different criterion variables are not clear. Observed changes in DIF identification rates may have been due to the subskills measured by the subtests or perhaps due to the short test length. More research in the area of matching criteria is needed.

In another study Tian, Pang, and Boss (1994a) examined the consistency of the MH and LR procedures across sample size and over replications. The same data set as above (Pang, Tian, & Boss, 1994) was used in this study with responses to the English test being analyzed. The 75 item English test is divided into two subtests: Usage/Mechanics (40 items) and Rhetorical Skills (35 items). Five levels of sample size (100/100, 250/250, 500/500, 1000/1000, and 2000/2000) were selected for comparisons between the reference group (females) and focal group (males) respectively. The MH and LR

procedures were first applied using total test score as the matching criterion then repeated using the subtest score as the matching criterion. Thirty replications were performed for each combination of sample size and matching criterion. Using the mean value of MH-Z from the 30 replications, items were classified as either Possible DIF, Probable DIF, or Definite DIF as outlined above (Pang, Tian, & Boss, 1994).

The results indicated that sample size had an effect on the MH and LR procedures. As the sample size increased, the power of the two procedures improved substantially. Using the subtest score as the criterion did not change the results of the MH and LR procedures much. The same items were identified as when the total test score was used as the criterion. In general, the agreement between the MH and LR procedures is very high across sample size and criterion variable. The LR procedure detected DIF somewhat more often than the MH procedure, but also had a slightly higher false-positive rate. In order for the procedures to perform reliably, the authors suggest sample sizes of at least 500 to 1000.

In a similar study Tian, Pang, and Boss (1994b) investigated the effects of sample size and criterion variable on the MH and LR procedures using data from a reading test. The same data set was used as above (Tian, Pang, & Boss, 1994a). The 40 item Reading Test is divided into two subtests: Reading Arts/Literature (20 items) and Reading Social Studies/Science (20 items). As above, males were the focal group and females were the reference group. Thirty replications were completed for each comparison of sample size (100/100, 250/250, 500/500, and 1000/1000) and criterion variable (total test score and subtest score). Classification of DIF items followed the guidelines above.

As in their previous study, sample size was observed to have a strong influence on the MH and LR procedures. As the sample size increased the power of both procedures to detect DIF increased. In general there was very high agreement between the MH and LR procedures in identifying DIF across sample sizes. However, as sample size decreased, the discrepancy between the two procedures became larger with the LR procedure having slightly larger detection rates.

The choice of criterion variable, total test score or subtest score, had a substantial effect on the identification of DIF items. Eleven items were identified as DIF using total test score as the criterion. Using subtest score as the criterion, five items fell into a different category of DIF and five new items were identified. Upon closer analysis of the Reading Test and its subtests, the authors suggest the multidimensionality of the test was probably the cause of these results.

The results support Tian, Pang, and Boss (1994a) in that the agreement between the MH and LR procedures in the identification of DIF items is very high across sample size and over replications.

Summary of Findings

In all of the studies examined above, the effect of sample size on the performance of the MH and LR procedures remained constant. An increase in sample size has been shown to increase DIF identification with both the MH procedure (Mazor et al., 1991; Miller & Oshima, 1992) and the LR procedure (Swaminathan & Rogers, 1990; Pang et al., 1994). Increasing the sample size increases the power of many statistical procedures; therefore,

the present finding was easily predicted. However, researchers continue to examine the effect of sample size in order to determine the lowest possible value that will adequately detect DIF. Tian, Pang, and Boss (1994a) suggest that a minimum sample size of 500 to 1000 be used for DIF identification. More information regarding sample size can be gained by examining the interaction of sample size with other variables such as percentage of DIF present, test length, or effect size.

Poorly discriminating items were least likely to be identified with the MH procedure (Mazor et al., 1991; Clauser et al., 1991a). Very difficult items were also least likely to be identified with the MH procedure (Mazor et al, 1991; Clauser et al., 1991a). In these cases, however, the items were made more difficult for the focal group. Perhaps, had the direction of DIF been reversed, different results may have occurred. Difference in p-values between the reference and focal groups may be more important than the differences in b.

By increasing the difference in the b values of the reference and focal groups, the amount of DIF for the given item is increased. Increasing the amount of DIF (effect size) has been shown to increase the rate of DIF detection with the MH procedure (Mazor et al., 1991; Clauser et al., 1991a; Miller & Oshima, 1992). In these studies, the amount of DIF was measured by the difference in the b values rather than the difference in the p-values between the reference and focal groups. Due to the nature of the normal curve, a difference in b of .6 has a different impact depending on the b value of the reference group and the item discrimination. For example, if the b value for the reference group was -.5, the b value for the focal group would be .1 corresponding to a difference in p-values of .061 and .117 for a values of .3 and .7, respectively. If the b value for the reference group

was .5 and therefore, 1.1 for the focal group, the difference in p-values for a values of .3 and .7 would be .056 and .103. Therefore, a given b-value difference has a different effect depending on the combination of a and b values. This difference in p-values has been examined in very few studies.

Miller and Oshima (1992) reported that as the percentage of DIF items in a test increased, the identification of these items by the MH procedure increased proportionately. This result was found to hold true for items of moderate DIF only. Conversely, Swaminathan and Rogers (1993) found that the detection rate of uniform DIF with the LR procedure actually increased from 70% to 76% when the percentage of DIF dropped from 15% to none other than the item of interest. The percentage of DIF did not affect the MH procedure (Swaminathan & Rogers, 1993).

In comparing the performance of the MH and LR procedures, researchers have generally found very high agreement rates (Swaminathan & Rogers, 1990; Swaminathan & Rogers, 1993; Brown, 1992; Pang et al., 1994). The LR procedure can identify both uniform and nonuniform DIF, whereas, the MH procedure can only identify nonuniform DIF when b departs considerably from 0. Tian, Pang, & Boss (1994a, 1994b) reported that the LR procedure produces slightly more false positives than the MH procedure.

Purpose of the Study

The purpose of this study was to examine the performance of the MH and LR procedures and their agreement in the identification of uniform DIF over replications with simulated data. Specifically, the following research questions were examined

1. Do b value difference and p-value difference effect DIF detection rates and which provides the best measure of effect size?
2. Does sample size effect DIF detection rates and what size of sample is needed for different kinds of items?
3. Does item discrimination effect DIF detection rates?
4. Does item difficulty effect DIF detection rates?
5. Does the percentage of DIF items in the test effect DIF detection rates?

CHAPTER III

METHODOLOGY

The methodology for the study is presented in this chapter. The chapter is divided into three sections: variables, data generation, and data analysis.

Variables

The performance of the MH and LR procedures was examined through manipulation of the following variables: sample size, percentage of DIF items, item discrimination (a), item difficulty (b), and effect size.

Sample Size

Using the revised DATAGEN program (Carlson, 1983), two data sets were generated to allow for comparisons between groups with equal ability distributions. These two examinee data sets comprised the Reference Group (RG) and the Focal Group (FG), each with a normal distribution of ability scores set to a mean of 0 and a standard deviation of 1. In few studies has there been an examination of the performance of the MH and LR procedures with sample sizes less than 500, yet test developers must often use small samples to validate their tests. For these reasons, five levels of sample size for the reference and focal groups were used: 100/100, 200/200, 400/400, 600/600, and 800/800.

Item Parameters

A 66 item test was generated (see Table 2). Item discrimination (a) was set such that the first 22 items had an a of .3 followed by 22 items with an a of .5, and 22 items with

an a of .7. These values represent biserial correlations (ρ) of approximately .29, .45, and .57, respectively (Crocker, & Algina, 1986, p. 351) and are, therefore, within a typical range for items in real life tests.

Table 2: The a parameters and b parameters for each item on the simulated 66 item test used in the study.

Item #	a	b RG	Item #	a	b RG	Item #	a	b RG
1	0.3	-1.9	23	0.5	-1.9	45	0.7	-1.9
2	0.3	-1.7	24	0.5	-1.7	46	0.7	-1.7
3	0.3	-1.5	25	0.5	-1.5	47	0.7	-1.5
4	0.3	-1.3	26	0.5	-1.3	48	0.7	-1.3
5	0.3	-1.1	27	0.5	-1.1	49	0.7	-1.1
6	0.3	-0.9	28	0.5	-0.9	50	0.7	-0.9
7	0.3	-0.7	29	0.5	-0.7	51	0.7	-0.7
8**	0.3	-0.5	30**	0.5	-0.5	52**	0.7	-0.5
9*	0.3	-0.5	31*	0.5	-0.5	53*	0.7	-0.5
10	0.3	-0.3	32	0.5	-0.3	54	0.7	-0.3
11	0.3	-0.1	33	0.5	-0.1	55	0.7	-0.1
12	0.3	0.1	34	0.5	0.1	56	0.7	0.1
13	0.3	0.3	35	0.5	0.3	57	0.7	0.3
14**	0.3	0.5	36**	0.5	0.5	58**	0.7	0.5
15*	0.3	0.5	37*	0.5	0.5	59*	0.7	0.5
16	0.3	0.7	38	0.5	0.7	60	0.7	0.7
17	0.3	0.9	39	0.5	0.9	61	0.7	0.9
18	0.3	1.1	40	0.5	1.1	62	0.7	1.1
19	0.3	1.3	41	0.5	1.3	63	0.7	1.3
20	0.3	1.5	42	0.5	1.5	64	0.7	1.5
21	0.3	1.7	43	0.5	1.7	65	0.7	1.7
22	0.3	1.9	44	0.5	1.9	66	0.7	1.9

* These items were targeted as DIF items in the study of 9% DIF items.
** These items were targeted as DIF items in the study of 18% DIF items.

Item difficulty (b) was set according to the following guidelines. Within each 22 item block of equal item discrimination, 20 items had a different item difficulty. The first item in each 22 item block had a b of -1.900 with each consecutive item having a b value

increased by 0.200. In each 22 item block there were two items with a b value of -0.500 and two items with a b value of 0.500. The pseudo guessing parameter (c) was held constant at .15.

Percentage of DIF Items

Two levels of percentage of DIF items in a test were examined. The percentage of DIF items in the test was altered by doubling the number of DIF items. Therefore, replications were completed for a test with approximately 9% DIF items (6 of 66) and for a test with approximately 18% DIF items (12 of 66). Doubling the percentage of DIF items involved changing six previously non-DIF items into DIF items. These six new DIF items were matched on all variables making them identical to the original six DIF items.

Effect Size

Effect size is a measure of the amount of DIF. The amount of uniform DIF was manipulated by increasing the item difficulty (b) for the focal group. This gave the reference group an advantage. The difference in b between the focal and reference group was used to induce DIF. Four levels of effect size (difference in b) were studied: .2, .4, .6, and .8. DIF items were those with b = -0.5 and +0.5 for the reference group. The variables studied are shown in Table 3.

Table 3: A summary of the variables that were manipulated in the study.

Level of Variables					
DIF Procedure	Sample Size RG / FG	% of DIF Items	Effect Size 24 unique p-value differences by combining levels of a, b, and b-difference		
			Item Discrimination a	Item Difficulty b	Effect Size Difference in b
MH	100 / 100	9%	0.3	-0.5	0.2
LR	200 / 200	18%	0.5	0.5	0.4
	400 / 400		0.7		0.6
	600 / 600				0.8
	800 / 800				

The differences in b values reflect differences in proportion correct (p-values) as shown in Table 4. P-value differences were used as a more accurate measure of effect size since the p-value of each item was contingent on the a and b values for the item. The following formula (Crocker & Algina, 1986) was used to compute the p-value for each item for focal and reference groups:

$$b_g = \frac{-\Phi^{-1}(p_g)}{\rho_g} \qquad (1)$$

where: b_g is the difficulty value for item g.

p_g is the proportion correct measure of item difficulty for item g.

$\Phi^{-1}(p_g)$ is the z-score that cuts off the area p_g to the left of z in the standard normal distribution.

ρ_g is the biserial correlation between the scores on item g and the latent trait scores; derived from a.

Since each p-value difference is a function of a, b, and difference in b between RG and FG, 24 distinct p-value differences were derived from the three levels of a, two levels of b, and four levels of b value difference. These 24 levels of p-value differences have been given a reference number from 1 to 24 in order from lowest to highest (see Table 4).

Data Generation

Responses to the 66 item test were generated based on the three-parameter IRT model, using the revised DATAGEN program (Carlson, 1983).

Table 4: Effect on p-value difference of b value difference, b value, and a value.

	Item #†	b	a	p	Z-score	p-value	p-value*	p-value difference RG - FG	p-value difference Reference #
Reference Group	8 & 9	-.5	.3	.29	.145	.5596	.6257		
	30 & 31	-.5	.5	.45	.225	.5910	.6523		
	52 & 53	-.5	.7	.57	.285	.6141	.6720		
	14 & 15	.5	.3	.29	-.145	.4404	.5243		
	36 & 37	.5	.5	.45	-.225	.4090	.4977		
	58 & 59	.5	.7	.57	-.285	.3859	.4780		
Focal Group b value difference .2	8 & 9	-.3	.3	.29	.087	.5359	.6055	.0202	2
	30 & 31	-.3	.5	.45	.135	.5557	.6223	.0300	4
	52 & 53	-.3	.7	.57	.171	.5675	.6324	.0396	7
	14 & 15	.7	.3	.29	-.203	.4207	.5076	.0167	1
	36 & 37	.7	.5	.45	-.345	.3745	.4683	.0294	3
	58 & 59	.7	.7	.57	-.399	.3446	.4429	.0351	5
Focal Group b value difference .4	8 & 9	-.1	.3	.29	.029	.5120	.5852	.0405	8
	30 & 31	-.1	.5	.45	.045	.5199	.5919	.0604	11
	52 & 53	-.1	.7	.57	.057	.5239	.5953	.0767	15
	14 & 15	.9	.3	.29	-.261	.3974	.4878	.0365	6
	36 & 37	.9	.5	.45	-.405	.3409	.4398	.0579	10
	58 & 59	.9	.7	.57	-.513	.3050	.4093	.0687	13
Focal Group b value difference .6	8 & 9	.1	.3	.29	-.029	.4880	.5648	.0609	12
	30 & 31	.1	.5	.45	-.045	.4801	.5581	.0942	18
	52 & 53	.1	.7	.57	-.057	.4761	.5547	.1173	21
	14 & 15	1.1	.3	.29	-.319	.3745	.4683	.0560	9
	36 & 37	1.1	.5	.45	-.4[illegible]	.3085	.4122	.0855	17
	58 & 59	1.1	.7	.57	-.627	.2643	.3747	.1033	19
Focal Group b value difference .8	8 & 9	.3	.3	.29	-.087	.4641	.5445	.0812	16
	30 & 31	.3	.5	.45	-.135	.4443	.5277	.1246	22
	52 & 53	.3	.7	.57	-.171	.4325	.5176	.1544	24
	14 & 15	1.3	.3	.29	-.377	.3520	.4492	.0751	14
	36 & 37	1.3	.5	.45	-.585	.2776	.3860	.1117	20
	58 & 59	1.3	.7	.57	-.741	.2296	.3452	.1328	23

† first item included for test with 6 biased items; both items included for test with 12 biased items

* corrected for guessing

Data Analysis

A program developed by Ackerman (1987) was used to estimate the MH indices, while a program developed by Spray (1991) was used to estimate the LR indices. Spray's program allows for the estimation of uniform and nonuniform DIF separately and therefore employs two separate chi-square indices, each with one degree of freedom. Detection rates were estimated for levels of significance set at .01 and .05.

A total of 100 replications was run for each combination of effect size and sample size and for each percentage of DIF items (9% and 18% DIF). Within each combination of effect size and sample size, all levels of item discrimination and item difficulty were automatically crossed by design. Regression analysis was performed predicting MH detection rates, LR detection rates, and Δ_{MH}. Sample size, percentage of DIF items, and item parameters were predictor variables for the models along with appropriate interactions. Results are also reported in percentage of correctly identified uniform DIF items over the 100 replications. The false positive rates for the MH and LR procedures are also reported.

CHAPTER IV

RESULTS AND DISCUSSION

The results of the study are reported and discussed in this chapter. The results of the regression analysis predicting MH detection rate, LR detection rate, and Δ_{MH} are reported in the first half of the chapter. A summary of the effect of the variables on the DIF detection rates and the outcome of the false positive analysis conclude the chapter.

Regression Analysis

In order to form conclusions about the impact of the variables on the performance of the MH and LR procedures a regression analysis was done. In regression terms, five models were generated predicting dependent variables as follows: MH detection rate at the .05 level ($MH_{.05}$), MH detection rate at the .01 level ($MH_{.01}$), LR detection rate at the .05 level ($LR_{.05}$), LR detection rate at the .01 level ($LR_{.01}$), and Δ_{MH}. Detection rate was the percentage of uniform DIF items identified by the procedure over 100 replications and Δ_{MH} was the mean Δ_{MH} over 100 replications. The predictor variables for the models were p-value difference (PDIFF), sample size (SSIZE), b value difference (BDIFF), percentage of DIF in the test (DIF), a value (A), and b value for the reference group (RGB). Various interactions between these variables were also examined.

Correlation coefficients (r) among all independent and dependent variables are shown in Table 5. The MH and LR detection rates were highly correlated at both levels

Table 5: Correlation matrix for the five dependent variables and six predictor variables.

	Pearson Product Moment Correlation (r)										
	MH p<.01	MH p<.05	LR p<.01	LR p<.05	Δ_{MH}	PDIFF P	SSIZE S	BDIFF B	DIF D	A A	RGB R
$MH_{.01}$	1.000										
$MH_{.05}$	.968*	1.000									
$LR_{.01}$	.994*	.978*	1.000								
$LR_{.05}$	.951*	.993*	.967*	1.000							
Δ_{MH}	-.717*	-.736*	-.758*	-.779*	1.000						
PDIFF	.716*	.739*	.759*	.784*	-.977*	1.000					
SSIZE	.530*	.560*	.499*	.515*	-.004	.000	1.000				
BDIFF	.577*	.619*	.615*	.661*	-.774*	.851*	.000	1.000			
DIF	-.066	-.066	-.065	-.068	.091	.000	.000	.000	1.000		
A	.365*	.369*	.383*	.387*	-.540*	.467*	.00	.000	.000	1.000	
RGB	-.070	-.072	-.080	-.083	.139*	-.102	.000	.000	.000	.000	1.000

* p<.01

of significance ($r_{MH.01*LR.01}$=.994) ($r_{MH.05*LR.05}$=.993). The detection rates were all highly negatively correlated with Δ_{MH}. MH detection rate increases with larger absolute values of $\Delta_{MH;}$ since the simulated DIF favoured the reference group, Δ_{MH} was negatively correlated with MH detection rate. LR detection rate was correlated with Δ_{MH} at a slightly higher rate than was MH detection rate. PDIFF was highly correlated with all the dependent variables, in particular Δ_{MH}, $r_{\Delta*P}$=-.977. SSIZE was correlated with the detection rates as was BDIFF and A. The independent variables that were significantly correlated with the detection rates of both procedures were consistently ranked in the same order, with PDIFF having the highest correlation with detection rates followed by BDIFF, SSIZE, and A. BDIFF and A were also correlated with Δ_{MH}. The percentage of DIF items on the test (DIF) and RGB were not significantly correlated with the detection rates.

In order to arrive at the most parsimonious regression models, an all-subsets regression analysis was performed for each dependent variable, using the six predictor variables as above and 15 two-way interactions derived from all-possible predictor variable interactions. Scatter plots of the residuals were completed for each model. None of these plots showed any violation of the assumptions of normality, linearity, independence, or equality of variances. Hence, the use of linear regression techniques was deemed appropriate for the data sets.

The r^2 values for the 6 one-variable models are shown in Table 6. PDIFF explained

Table 6: Proportion of variance (r^2) accounted for by the 6 one-variable models predicting MH detection rate, LR detection rate, and Δ_{MH}.

One Variable Models		r^2 for Model				
		$MH_{.05}$	$MH_{.01}$	$LR_{.05}$	$LR_{.01}$	Δ_{MH}
PDIFF	P	.547*	.513*	.614*	.575*	.954*
SSIZE	S	.313*	.281*	.266*	.249*	.000
BDIFF	B	.383*	.333*	.436*	.378*	.598*
DIF	D	.004	.004	.005	.004	.008
A	A	.136*	.133*	.150*	.147*	.292*
RGB	R	.005	.005	.007	.006	.019*

* $p<.01$

the largest amount of variance in all five models; $MH_{.05}$ (r^2_P=.547), $MH_{.01}$ (r^2_P=.513), $LR_{.05}$ (r^2_P=.614), $MH_{.01}$ (r^2_P=.575), and Δ_{MH} (r^2_P=.954). These values were quite high especially for Δ_{MH}.

The six predictor variables resulted in 15 possible two-way interactions. These interactions were tested using variable-by-variable tests as suggested by Darlington (1990). When using variable-by-variable tests, one performs one test for all the

interactions involving a single regressor and then corrects for the number of tests performed. Six predictor variables resulted in six tests of interactions for each model. Using the Bonferroni method, these tests were adjusted accordingly. R^2 changes between the model with interaction terms and the model without interaction terms (P+S+B+D+A+R) were tested (see Table 7).

Table 7: R^2 values for the variable-by-variable tests of all possible two-way interactions predicting $MH_{.05}$, $MH_{.01}$, $LR_{.05}$, $LR_{.01}$, and Δ_{MH}.

Model	$MH_{.05}$	$LR_{.05}$	$MH_{.01}$	$LR_{.01}$	Δ_{MH}
P+S+B+D+A+R	.8651	.8851	.8034	.8332	.9757
P+S+B+D+A+R+P*S+P*B+P*D+P*A+P*R	.9213* F=49.84	.9238* F=35.45	.9438* F=174.38	.9406* F=126.20	.9839* F=35.65
P+S+B+D+A+R+S*B+S*D+S*A+S*R+P*S	.9266* F=58.48	.9297* F=44.28	.9429* F=173.26	.9405* F=125.87	.9759 F=.579
P+S+B+D+A+R+B*D+B*A+B*R+P*B+S*B	.9169* F=45.51	.9250* F=37.13	.9088* F=80.67	.9177* F=71.67	.9805* F=17.18
P+S+B+D+A+R+D*A+D*R+P*D+S*D+B*D	.8661 F=.521	.8858 F=.428	.8055 F=.754	.8351 F=.804	.9790* F=10.97
P+S+B+D+A+R+A*R+P*A+S*A+B*A+D*A	.8765* F=6.44	.8918* F=4.32	.8365* F=14.12	.8573* F=11.79	.9809* F=19.00
P+S+B+D+A+R+P*R+S*R+B*R+D*R+A*R	.8658 F=.364	.8856 F=.305	.8038 F=.142	.8335 F=.126	.9970 F=3.95

* $F_{6,346}$ ($p<.0017$)

Results of the variable-by-variable tests indicated that interactions involving DIF and RGB were not significant in predicting detection rates. Interactions involving SSIZE and RGB were not significant in predicting Δ_{MH}. Interactions involving other variables were significant. Simple interaction tests were completed to test these specific interactions to determine which to include in the model (see Table 8).

Table 8: R^2 values for simple interaction tests involving PDIFF, SSIZE, BDIFF, and A for the models predicting DIF detection rate and simple interaction tests involving PDIFF, BDIFF, A, and DIF for the model predicting Δ_{MH}.

Model	$MH_{.05}$	$LR_{.05}$	$MH_{.01}$	$LR_{.01}$	Δ_{MH}
P + S P + S + P*S	.8600 .9138 F=222.19*	.8795 .9141 F=143.39*	.7934 .9309 F=708.39*	.8241 .9291 F=527.22*	
P + B P + B + P*B	.5470 .5491 F=1.66	.6141 .6185 F=4.11	.5163 .5188 F=1.85	.5780 .5807 F=2.46	.9668 .9688 F=22.82*
P + A P + A + P*A	.5473 .5474 F=.079	.6145 .6151 F=.555	.5137 .5185 F=3.55	.5764 .5805 F=3.47	.9636 .9683 F=52.78*
S + B S + B + S*B	.6965 .7459 F=69.21*	.7020 .7378 F=48.61*	.6143 .7171 F=129.36*	.6270 .7092 F=100.63*	
S + A S + A + S*A	.4496 .4606 F=7.26	.4152 .4214 F=3.81	.4143 .4459 F=20.30*	.3957 .4185 F=13.96*	
B + A B + A + B*A	.5192 .5373 F=13.93*	.5861 .6025 F=14.69*	.4667 .5092 F=30.63*	.5252 .5690 F=36.18*	.8901 .9412 F=309.38*
P + D P + D + P*D					.9628 .9660 F=33.51*
B + D B + D + B*D					.6067 .6088 F=1.91
D + A D + A + D*A					.3002 .3011 F=.458

* $F_{1,344}$ ($p<.01$)

PDIFF*SSIZE, SSIZE*BDIFF, and BDIFF*A were significant interactions for the four models predicting DIF detection rates. SSIZE*A was significant for $MH_{.01}$ and $LR_{.01}$ but not for $MH_{.05}$ and $LR_{.05}$. PDIFF*BDIFF, PDIFF*A, BDIFF*A, and PDIFF*DIF were significant interactions for the model predicting Δ_{MH}. At this point it becomes clear that PDIFF, SSIZE, and PDIFF*SSIZE seem to explain a large portion of the variance in detection rates; however, the analysis continued with an all-subsets regression using the significant

predictor variables and significant interactions, in order to generate the most parsimonious regression model for each dependent variable (see Table 9). The difference in the number

Table 9: R^2 values for the best one-, two-, three-, and four-variable models, along with the omnibus model, predicting $MH_{.05}$, $MH_{.01}$, $LR_{.05}$, $LR_{.01}$, and Δ_{MH}.

# of Variables in Model	Model Predicting $MH_{.05}$	R^2	Model Predicting $MH_{.01}$	R^2	Model Predicting Δ_{MH}	R^2
Best One Variable Model	P	.547	P	.513	P	.955
Best Two Variable Model	P + S	.860	P + S	.793	P + B	.967
Best Three Variable Model†	P + S + P*S P + S + A	.914 .861	P + S + P*S P + S + B	.931 .797	P + B + D	.975
Best Four Variable Model†	P + S + P*S + A P + S + A + B	.915 .861	P + S + P*S + B P + S + B + A	.935 .798	P + B + D + P*D P + B + D + A	.978 .975
Omnibus Model†	7 variables 4 variables	.917 .861	8 variables 4 variables	.938 .798	8 variables 4 variables	.983 .975
# of Variables in Model	Model Predicting $LR_{.05}$	R^2	Model Predicting $LR_{.01}$	R^2		
Best One Variable Model	P	.614	P	.575		
Best Two Variable Model	P + S	.880	P + S	.824		
Best Three Variable Model†	P + S + P*S P + S + A	.914 .880	P + S + P*S P + S + B	.929 .827		
Best Four Variable Model†	P + S + P*S + A P + S + A + B	.915 .880	P + S + P*S + B P + S + B + A	.933 .829		
Omnibus Model†	7 variables 4 variables	.919 .880	8 variables 4 variables	.936 .829		

† The second model is the best model without allowing for any interaction terms.

of variables among the omnibus models is due to the lower number of significant interactions that were used to generate the model at the .05 level of significance. For comparison sake, each of the best models is juxtaposed with the best model without interaction terms where applicable.

MH and LR Detection Rates

The omnibus models, created by the four significant predictor variables (PDIFF, SSIZE, BDIFF, and A) and all significant two-way interactions, explained a high percentage of variance in MH and LR detection rates at both levels of significance. However, the most parsimonious model explaining detection rates was PDIFF + SSIZE + PDIFF*SSIZE which explained 5% more variance in $MH_{.05}$ and 3% more variance in $LR_{.05}$ than the best two-variable model (P+S); at the .01 level, 14% more variance was explained in $MH_{.01}$ and 11% in $LR_{.01}$. For both procedures at each level of significance, the addition of a fourth term to the model improved the variance by less than .5%. The model for $MH_{.05}$ is examined first followed by the other three models.

The complete model for $MH_{.05}$ with b-weights was as follows:

$$MH_{.05} = -12.45 + 303.91 \text{ PDIFF} + .014 \text{ SSIZE} + .767 \text{ PDIFF*SSIZE}$$

The corresponding standard errors of the b-weights were 2.024, 25.311, .0041, and .0515, The standard error of estimate was 9.32. This model explained 91% of the variance in MH at the .05 detection rate.

The model can be explained by examining the b-weights beginning with the scaling differences in the variables. The b-weight for PDIFF was very large as PDIFF alone explained 55% of the variance. However, the actual p-value differences were quite small, in the range of .0167 and .1544, which artificially inflated the magnitude of the b-weight. Although the b-weights for SSIZE and PDIFF*SSIZE were much lower, their relative weight is deceiving since sample size, in effect, acts as a multiplier by a factor of 100.

The interaction of PDIFF*SSIZE added 5% to the amount of variance explained by the model. As SSIZE increased the effect of PDIFF on the MH detection rate increased as well. P-value difference had a larger effect at higher sample sizes. The reverse is also true; sample size had a larger effect at higher p-value differences. The conditional effect of PDIFF on $MH_{.05}$ when sample size is 100 is 380.61. Which means for each increase in p-value difference of .0500, $MH_{.05}$ increases by 19%. When sample size is increased to 800, the conditional effect of PDIFF becomes 917.51. At sample size 800, each increase in p-value difference of .0500 increases $MH_{.05}$ by 46%. This reflects the data where p-values greater than .1000 were detected approximately 100% at sample size 800. The effect of PDIFF is larger at larger sample sizes. At sample sizes above 400, the effect of the interaction term has a greater effect on $MH_{.05}$ than the effect of PDIFF alone. For each increase in sample size of 100, the conditional effect of PDIFF increases by 76.7.

When PDIFF is low (.0500), the conditional effect of sample size is .0524. Which means for each increase in sample size of 100, $MH_{.05}$ increases by 5.24%. However, when PDIFF is high (.1500), the conditional effect of sample size more than doubles to .1291. At a high level of PDIFF, each increase in sample size of 100 increases $MH_{.05}$ by 12.91%. Therefore, as mentioned above, sample size has a larger effect on $MH_{.05}$ at higher p-value differences.

This model can be used to predict the percentage of uniform DIF that would go undetected by the MH procedure at significance level of .05. In order for the model to be used for prediction, conditions similar to those in this study must hold true, i.e. a 66 item test. Assuming similar conditions, if a test developer with 300 examinees wanted to know

what percent a relatively high level of DIF (p-value difference = .100) would go undetected with the $MH_{.05}$ procedure, the regression model would be as follows:

$$MH_{.05} = -12.45 + 303.91(.100) + .014(300) + .767(.100)(300)$$

Solving for the MH detection rate (45%) and subtracting from 100 shows that the test developer can expect 55% +/- 20% (95% confidence interval) of the DIF items to go undetected at $p<.05$. Another example would be the test developer who wanted to predict the sample size needed to detect a relatively low level of DIF (p-value difference = .050) with 95% accuracy; the model would resemble the following:

$$95 = -12.45 + 303.91(.050) + .014(SSIZE) + .767(.050)(SSIZE)$$

Solving for sample size would mean that this particular test developer would require a sample size of at least 1762. If the test developer would settle for a DIF detection rate of 80%, only 1476 examinees would be required. At the very least, the model demonstrated that p-value difference and sample size are good predictors of $MH_{.05}$ detection rates and that p-value difference is an improvement over the more traditional measure of effect size, b value difference.

A very similar model was found for $LR_{.05}$:

$$LR_{.05} = -12.59 + 399.70\ PDIFF + .019\ SSIZE + .607\ PDIFF*SSIZE$$

Corresponding standard errors of the b-weights were 2.00, 24.92, .004, and .051. The standard error of estimate was 9.18.

Similar conclusions about the relative size of the b-weights can be made in regards to this model. An increase in PDIFF will increase $LR_{.05}$ detection rate. Increasing sample size will augment the effect of p-value difference. For each increase in sample size of 100,

the effect of PDIFF will increase by 60.7. At low (.0500) and high (.1500) PDIFF, the conditional effect of sample size is .0494 and .1101, respectively. Therefore, even though the b-weight for SSIZE was higher in the $LR_{.05}$ model than in the $MH_{.05}$ model, sample size had more of an overall effect on $MH_{.05}$ than $LR_{.05}$.

As with the previous model, this model can be used for prediction purposes with a 95% confidence interval of +/- 20% around $LR_{.05}$, provided conditions similar to those in this study are met.

The b-weights for the models predicting DIF detection rate at the .01 level of significance were slightly different than the models described above:

$$MH_{.01} = -6.55 + 85.68\ PDIFF - .0206\ SSIZE + 1.150\ PDIFF*SSIZE$$

$$LR_{.01} = -9.59 + 183.46\ PDIFF - .0141\ SSIZE + 1.022\ PDIFF*SSIZE$$

Standard errors of the b-weights for $MH_{.01}$ were 1.708, 21.25, .0035, and .0432; with a standard error of estimate of 7.82. For $LR_{.01}$ the standard errors were 1.760, 21.90, 0036, 0445; with a standard error of estimate of 8.063.

The most obvious differences are the lower b-weight for PDIFF, the negative b-weight for SSIZE, and the higher b-weight for PDIFF*SSIZE. One effect of these differences is that the conditional effect of PDIFF on $MH_{.05}$ is larger than on $MH_{.01}$ at small sample sizes; however, as SSIZE increases, the conditional effect of PDIFF on $MH_{.01}$ becomes larger than on $MH_{.05}$. For example, the conditional effect of PDIFF on $MH_{.01}$ when SSIZE is 100 is 200.68 compared to 380.61 on $MH_{.05}$. At sample size 800, the conditional effect of PDIFF on $MH_{.01}$ is 1005.68 compared to only 917.51 on $MH_{.05}$. This suggests that sample size has more of an influence on p-value difference at the .01 level of significance

levels than at .05. Perhaps a ceiling effect on DIF detection rate existed at the .05 level of significance as approximately 10% of the detection rates were 100% compared to only 2.5% at the .01 level of significance. These high detection rates were observed at the larger sample sizes. Therefore, sample size may have had more of an effect at the .01 level of significance as there was more variability in the detection rates.

When PDIFF is low, .0500, the conditional effect of SSIZE is .0369 on $MH_{.01}$ compared to .0524 on $MH_{.05}$. As PDIFF increases (.1500), the conditional effect of SSIZE on DIF detection rate increases to the point that the effect on $MH_{.01}$ (.1519) is larger than on $MH_{.05}$ (.1291). It would appear that p-value difference has more of an influence on sample size at lower significance levels. Furthermore, when PDIFF and SSIZE are the same for both models, PDIFF*SSIZE has more of an effect at the .01 level.

The negative b-weight for SSIZE must not be misinterpreted to imply that as sample size increases, DIF detection rates decrease. When p-value difference is greater than .0180, an increase in sample size will increase $MH_{.01}$; for p-value differences greater than .0138, an increase in sample size will increase $LR_{.01}$. For smaller p-value differences, the models predict that as sample size increases, DIF detection rate will decrease. This is due simply to a chance occurrence that can be found in the data used to generate the models at the .01 level of significance. In the study only one item had a p-value difference less than these limits (p-diff. reference #1). The original data for this item at the .01 level of significance is shown in Table 22 (Appendix A). In fact, the detection rates for this item reached values of 4% and 6% with the MH and LR procedures, respectively, at sample size 600 and then dropped to 1% and 2% at sample size 800. Therefore, with this item (p-

diff=.0167), an increase in sample size was followed by a decrease in the DIF detection rate. This item was at chance rates throughout the study. This chance event and one or two others like it at low p-value difference levels (Table 22, Appendix A) probably caused the negative b-weight in the models at the .01 level since the effect of sample size is only reversed at such a low level of p-value difference. For all general purposes, these models predict that an increase in sample size will increase detection rates.

As expected, lower detection rates were predicted using the .01 level models as compared to the .05 models. In the earlier example, with a p-value difference of .100 and a sample size of 300, the $MH_{.05}$ detection rate was 45% +/- 20%. At the .01 level, the detection rate would be 30% +/- 17%. The standard error of estimate is lower for the .01 level models which means that detection rates can be predicted with greater accuracy.

In summary, when predicting MH and LR detection rates, PDIFF + SSIZE + PDIFF*SSIZE seems to be the most effective model. The best model predicting Δ_{MH} is explained in the next section.

$\underline{\Delta_{MH}}$

Clearly, p-value difference (r^2_P=.954) had a very strong influence on Δ_{MH} explaining 35% more variance in Δ_{MH} than did b value difference (r^2_B=.598). The omnibus model, created by the six predictor variables and four significant two-way interactions, explained a high percentage of variance in Δ_{MH} (R^2=.983). The best one-, two-, three-, and four-variable models are shown in Table 9. PDIFF was by far the best predictor of Δ_{MH}; yet because of the degrees of freedom (df=1,358), results from the F tests of R^2 changes as additional terms were added to the model were significant. Even though additional terms

significantly increased the R^2 of the model, this increase was only by approximately 1% at each step. After examining the b-weights of the two- and three- variable models it was obvious that the additional terms were theoretically insignificant albeit statistically significant. For these reasons, the most parsimonious model predicting Δ_{MH} was deemed to be the one-variable model, explaining 95.5% of the variance:

$$\Delta_{MH} = .025 - 9.789 \text{ PDIFF}$$

The standard errors of the b-weights were .009 and .113. The standard error of estimate was .080.

In the study, the DIF items were made easier for the reference group. As a result, the Δ_{MH} values were negative. According to the model, as PDIFF increases, Δ_{MH} becomes more negative, making the item more likely to be detected by the MH procedure.

Summaries of the effects of the individual variables on DIF detection rate are presented in the following section.

DIF Detection Rates

In general, a wide range of DIF identification rates was produced with both procedures ranging from 0% to 100%. Definite trends were observed in the data which are discussed in the following paragraphs according to the variable of interest. Summary tables are presented throughout the discussion; however, more detailed tables are also referenced and can be found in Appendix A.

DIF Identification Procedure and Sample Size

The agreement between the LR and MH procedures was high over all conditions as shown in Table 10. The LR procedure identified the DIF items as well as or marginally better than the MH procedure in 98% of the cases. The discrepancy between the two procedures is greater at the .05 level of significance than at .01. Supporting data for the

Table 10: Summary of percent detection rates for uniform DIF over 100 replications for the MH and LR procedures averaged across levels of sample size.

Sample Size	$MH_{.05}$	$LR_{.05}$	$MH_{.01}$	$LR_{.01}$
100 Per Group	12.16	18.74	4.25	7.58
200 Per Group	24.25	29.92	12.11	15.49
400 Per Group	43.26	46.39	26.59	30.54
600 Per Group	53.78	56.57	39.23	41.2
800 Per Group	61.5	63.5	47.44	48.9
Average	38.99	43.02	25.92	28.74

9% and 18% DIF tests at the .05 and .01 level of significance are shown by sample size and p-value difference in Tables 18, 19, 20, and 21 (Appendix A).

Using real data, Tian, Pang, and Boss (1994a) found similar results, in that the LR procedure produced slightly higher DIF identification frequencies than the MH procedure. On the contrary, Rogers and Swaminathan (1993) found that the MH procedure had a slight advantage over the LR procedure in detecting uniform DIF. Their results may have been due to a two-stage process used with the MH and LR procedures or perhaps to a chi-

square test with 2 df. The two-stage process was not used in the present study and the chi-square test used had 1 df.

Under all conditions, sample size had a positive relationship to the number of times an item was correctly identified with the MH and LR procedures (see Table 10). Increasing the sample size improved the power of the MH and LR chi-square statistics. This finding is supported by Tian, Pang, and Boss (1994b) and Rogers and Swaminathan (1993) who also found that an increase in sample size increased the MH and LR detection rates. Mazor, Clauser, and Hambleton (1991) came to the same conclusion with the MH procedure.

At lower sample sizes, the difference in the identification rates of the two procedures was higher than at larger sample sizes. For example, the DIF identification rates for the $LR_{.05}$ procedure were on average 2% higher per item than those for the $MH_{.05}$ procedure at sample size 800. At sample size 400, the discrepancy between the two procedures increased to 3% per item and at sample size 100, the $LR_{.05}$ procedure outperformed the $MH_{.05}$ procedure by 7%. This effect was also demonstrated at the .01 level of significance, but not to the same degree.

Recall that the regression models predicted that increasing sample size would increase the effect of p-value difference on the $MH_{.05}$ and $LR_{.05}$ detection rates. Moreover, for each increase in sample size of 100, the effect of p-value difference would increase by 76.7 on the $MH_{.05}$ detection rate, but by only 60.7 on the $LR_{.05}$ detection rate. According to the regression models and the present observations, increasing the sample size had more of an effect on the MH procedure than on the LR procedure. Tian, Pang, and Boss

(1994a) also observed the increased discrepancy between the two procedures at lower sample sizes.

Items with large amounts of DIF (p-diff>.100) were identified at least 70% of the time with sample sizes as low as 400 (p<.05). Items with large amounts of DIF were only identified, on average, 44% of the time with sample sizes 100 (p<.05). At sample size 100, items with low DIF (p-diff<.050) were not identified any more frequently than were items with no DIF at all. One would probably not want to use a sample size lower than 400 when using the MH or LR procedure.

a Value

Each level of a (.3, .5, and .7) included eight items in different combinations of b value difference and b value. Results in this section are based on averages of each group of eight items at each level of a. DIF detection rates increased with increasing a values (see Table 11). Supporting data for the 9% and 18% DIF tests at the .05 and .01 level of

Table 11: Summary of percent detection rates for uniform DIF over 100 replications for the MH and LR procedures across levels of a.

a value	$MH_{.05}$	$LR_{.05}$	$MH_{.01}$	$LR_{.01}$
.3	23.55	27.12	11.87	13.61
.5	40.70	45.01	27.04	30.12
.7	52.48	56.81	38.85	42.49

significance are shown by sample size across levels of a value, b value, and b value difference in Tables 22, 23, 24, and 25 (Appendix A).

As the a value increased, the DIF detection rates for both the MH and LR procedure increased as well. Because of the nature of the p-value formula, larger a values resulted in larger p-value differences. If b is held constant, items with higher a values will have larger p-value differences. Hence, on average, items with an a value of .7 were detected more frequently than items with an a value of .5.

According to Formula 1, a and b values combine together to produce p-values and then p-value differences. Therefore, since b is not always constant, one cannot assume that items with large a values will be identified more frequently than items with lower a values. In this study, DIF detection rates were reported using p-value difference since it is a quantitative measure of the combination of a, b, and b value difference. Some authors, however, have explained the combination of a and b in a descriptive manner and reported the DIF detection rates as such.

In order to describe the interplay between a and b values, Rogers and Swaminathan (1993) used a descriptive method. The authors studied four different types of items, with four sample items of each type. The four types of items were described as follows with typical item parameters in brackets: low b, high a (-1.83, 1.2); moderate b, low a (-.34, .6); moderate b, high a (-.34, 1.2); and high b, high a (1.18, 1.2). Using this descriptive model, the two levels of b used in this study (-.5 and .5) would be considered moderate; the three levels of a (.3, .5, and .7) would be considered low. According to Rogers and Swaminathan (1993), other than the size of the DIF, the largest effect observed for both procedures was due to the type of item. Over 20 replications, items of moderate b and high a had the highest detection rates for both procedures. Items of moderate b and low

a had the lowest detection rates for the LR procedure; whereas, items of high b and high a had the lowest detection rates for the MH procedure.

The method of describing the interplay between b and a used by Rogers and Swaminathan (1993) can be translated into the more quantitative method of p-value difference used in this study. If one converts the above b and a values into p-value differences, for a b value difference of .4, the descriptive and quantitative method can be compared as follows: low b, high a (p-diff=.0440); moderate b, low a (p-diff=.0636); moderate b, high a (p-diff=.0980); and high b, high a (p-diff=.0568). Items with moderate b and high a values had the highest p-value difference and, therefore, had higher identification rates than the other items with both procedures as reported by Rogers and Swaminathan (1993). Items with low b and high a values had the lowest p-value difference and, therefore, one would have expected these items to have the lowest identification rates with both procedures. However, this was not the case. Since the differences in p-value difference among the other three types of items were quite small, the identification rates were quite close for these three types of items. DIF identification rates differed by less than 5%. Results may have been due to the limited number of replications used. Perhaps more replications might have resulted in more accurate DIF detection rates. Guessing may have also effected the identification rates, especially at the moderate and high levels of b.

b Value

Each level of b (-.5 and .5) included twelve items in different combinations of b value difference and a value. Results in this section are based on averages of each group

of twelve items at each level of b (see Table 12). Both procedures identified items with a b value of -.5 more frequently than items with a b of .5. These findings are due to the increased p-value differences associated with items with a b value of -.5. This result can be explained by going back to the p-value formula (Formula 1).

Table 12: Summary of percent detection rates for uniform DIF over 100 replications for the MH and LR procedures across levels of b.

b value	$MH_{.05}$	$LR_{.05}$	$MH_{.01}$	$LR_{.01}$
-.5	41.22	45.54	28.09	31.10
.5	36.59	39.42	23.76	26.39

In order to calculate a p-value difference, a p-value for the reference and focal group must be calculated from a z-score. According to the formula, the z-score is the negative product of b and ρ (derived from a). Since ρ is always positive, the z-score has the opposite sign from the b value. Therefore, an item with a negative b value will have a positive z-score. If the item favours the reference group, the b value of the focal group is increased by the appropriate b value difference; hence, the z-score of the focal group is decreased (negative product). The difference between these z-scores is the p-value difference. When comparing areas under the normal curve, a larger area will result from the difference between a positive z-score and a lesser z-score than the same difference between a negative z-score of same magnitude and a lesser z-score. The following conclusions can be drawn from the p-value formula providing both items have the same

a value, the same b value difference, and the same absolute b value. If the DIF favours the reference group, the item with the negative b value will have a larger p-value difference and, therefore, be detected more frequently than the item with the positive b value. If the item favours the focal group, then the item with the positive b value will be detected more frequently. In the present study, since the simulated DIF favoured the reference group, the item with the negative b value was consistently detected more frequently than the matched item with the positive b value.

According to Formula 1, it would seem that the effect b, when a is constant, cannot be generalized, even if the direction of the DIF is constant as well. In other words, one cannot state that lower levels of b will have higher DIF detection rates if the a value is constant. An example is provided in Table 13.

Table 13: An example illustrating that the effect of b value on DIF detection rates cannot be generalized across different levels of a.

	b	a	p	z-score	p-value	p-value*	p-value difference RG-FG
Reference Group	-.9	.5	.45	.405	.6591	.7102	
	-.3	.5	.45	.135	.5557	.6223	
	.3	.5	.45	-.135	.4443	.5277	
	.9	.5	.45	-.405	.3409	.4398	
Focal Group	-.5	.5	.45	.225	.5910	.6524	.0578
	.1	.5	.45	-.045	.4801	.5581	.0642
b-value difference .4	.7	.5	.45	-.315	.3745	.4683	.0594
	1.3	.5	.45	-.585	.2776	.3860	.0538

* corrected for guessing

In the example, a was held constant at .5 and b ranged from -.9 to +.9. As b increased, there was no observable trend in p-value difference. Assuming that items with higher p-value differences will be detected more frequently, the item with b of -.9 would be detected more frequently than the item with b of .9, yet less frequently than the item with b of .3. However, since the DIF favoured the reference group in the example, items with negative b values had higher p-value differences than their corresponding positive match. This supports the results explained above.

Clauser, Mazor, and Hambleton (1991a) observed the effect of b values on DIF detection with the MH procedure. Five levels of b were studied: -2.5, -1.0, 0.0, 1.0, 2.5. Four levels of a and b value difference were completely crossed with the levels of b. Simulated DIF favoured the reference group. According to the results from this study, items with a b value of -2.5 should be detected more frequently than items with a b value of 2.5. The results of their study support these results; detection rates for items with b values of -2.5 and 2.5 were 69% and 13%, respectively. Following the same trend, detection rates for items with b values of -1.0 and 1.0 were 81% and 69% respectively.

b Value Difference

Each level of b value difference (.2, .4, .6, and .8) included six items in different combinations of a and b. Results in this section are based on averages of each group of six items at each level of b value difference. Items with larger b value differences were detected more frequently by both procedures (see Table 14).

These findings are not surprising as one would expect an increase in effect size to increase detection rates. Similar results were found with the MH procedure by Mazor et.

Table 14: Summary of percent detection rates for uniform DIF over 100 replications for the MH and LR procedures across levels of b value difference.

b value difference	$MH_{.05}$	$LR_{.05}$	$MH_{.01}$	$LR_{.01}$
.2	11.51	13.66	4	4.77
.4	30.35	34.59	16.01	18.35
.6	49.93	54.66	34.06	37.57
.8	63.84	69.02	49.63	54.27

al. (1991), Clauser et al., (1991a), and Miller and Oshima (1992). The intent of the author is not to discount the relationship between b value difference and DIF detection rates, since a strong positive relationship has been shown to exist. However, as observed with the regression analysis and as discussed in the next section, the author suggests that b value difference is a rather crude measure of effect size. P-value difference can lead to a more realistic measure of effect size and, therefore, better prediction of DIF detection.

p-value Difference

Increasing the effect size, p-value difference, also improved the rate of DIF identification with both procedures across all conditions (see Table 15). This finding is supported by the regression models.

In the 9% DIF test, there were six items within each of the four levels of b value difference (.2, .4, .6, and .8). Each of these six items was a unique combination of one of two b values for RG (-.5, .5) and one of three a values (.3, .5, and .7). Within each of the four levels of b-difference, these six items, ranked according to increasing p-value differences, were as follows (RGb,a): (.5, .3), (-.5, .3), (.5, .5), (-.5, .5), (.5, .7), (-.5, .7).

Table 15: Summary of percent detection rates for uniform DIF over 100 replications for the MH and LR procedures across levels of p-value difference.

p-value difference*	$MH_{.05}$	$LR_{.05}$	$MH_{.01}$	$LR_{.01}$
p-diff < .050	12.49	14.78	4.34	5.24
.050 ≤ p-diff < .075	32.72	36.63	16.75	18.86
.075 ≤ p-diff < .100	48.07	53.79	30.40	34.29
p-diff ≥ .100	71.98	77.03	58.62	63.69

* Ranges of p-value difference were selected in order to have a relatively equal number of items (8, 5, 5, and 6, respectively) within each range.

In the 18% DIF test, the items were ranked in the same order; however, there were two items for each combination of b and a. For both tests, at both levels of significance ($p<.05$ and $p<.01$), across all sample sizes, and for both procedures, the DIF identification rates within these levels of b value difference were ranked in the same order from low to high (ie. item (.5, .3) identified the least, to item (-.5, .7)) with few exceptions.

Different combinations of a, b and b-differences resulted in different p-value differences. Using the formula for obtaining the p-value of an item (Formula 1), it can be noted that the z-score, used to cut off the area p to the left of z, is actually the negative product of b and a (since p is a function of a). Understanding the normal curve, allows one to make predictions about DIF detection rates. For instance, in the present study, an item with b (.5) and a (.7) resulted in a z-score of -.285 and a corresponding p-value of .3859. When compared with a b-difference of .6 (b = 1.1), the resulting p-value difference was .1033. In this case, the reference group had a 10% better chance of getting the item correct than the focal group. Using the same parameters, changing only the RG b to -.5 (z-score .285) and, hence, the FG b to .1, gave the reference group a 12% better chance

of getting the item correct because the p-value difference between the two groups was increased with the new z-score. Therefore, it has been observed that whichever combination of a, b, and b-difference causes the greatest difference in the proportion-correct area under the curve, results in a higher DIF detection rate.

The p-value difference proved to be a more realistic measure of effect size than the b value difference alone. When the items were ordered according to p-value differences, the increasing detection rate was better explained. Some items at the .4 level of b value difference were identified more frequently than some items at the .6 level because the combination of b and a values resulted in larger p-value differences. There was considerable overlap at the .4 and .6 levels of b value difference across all sample sizes with both procedures. One item at the .4 level had a correspondingly larger p-value difference than an item at the .8 level of b value difference and, as such, was identified more frequently. These large ranges of DIF detection rates can not accurately be explained using b value differences alone.

In summary, this trend of increasing p-value differences accompanied by increasing DIF detection rates can be related back to the variables in the original formula. Particular combinations of a, b, and b value difference undoubtedly result in larger p-value differences for items. These items, in turn, were identified more frequently with the MH and LR procedures. In this study, when comparing DIF detection rates of items with the same RG b and b value difference, the item with the largest a value was detected more frequently. When comparing DIF detection rates of items with the same a and b value difference, the item with the lower b value was detected more frequently. Finally, when

comparing items with the same b and a values the item with the largest b value difference was detected more frequently.

According to the definition of DIF, an item is considered to function differently if both groups do not have the same probability of getting the item correct. It stands to reason, therefore, that a more precise measure of the probability, would produce a more accurate measure of DIF. Using the actual p-value difference is clearly a more precise measure of the amount of DIF in a given item than is using the b value difference alone. Having a more accurate measure of the amount of DIF, allows for a more accurate explanation of the DIF detection rates.

Rogers and Swaminathan (1993) studied the effect of the amount of DIF on the performance of the MH and LR procedures and came to the same conclusion, that the larger the effect size, the greater the probability of detection. The size of DIF was quantified in terms of the area between the generating item response functions using the formula given by Raju (1988). Clauser, Mazor, and Hambleton (1991a) and Miller and Oshima (1992) manipulated the size of DIF by increasing the b value difference between the RG and FG. Both sets of authors observed that by increasing the size of DIF, the chance of the item being detected by the MH procedure increased dramatically. However, Clauser, Mazor, and Hambleton (1991a) concede that this pattern of increasing detection rates was highly influenced by the other variables examined, namely a and b. The notion that perhaps these variables contributed to another value, for example, a p-value difference, was not discussed.

In another study, Mazor, Clauser, and Hambleton (1991) made mention of p-value differences. They reported increasing DIF detection rates with the MH procedure with increasing b value differences. Included in their paper was a table of p-value differences which came from various combinations of b, a, and b value difference. Smallest p-value differences detected and largest p-value differences missed were recorded at each level of sample size. In their study only one or two replications were completed for each item. However, the general trend of increasing DIF detection rates with increasing p-value differences was observed. The authors did not discuss whether p-value difference or b value difference resulted in a more realistic measure of effect size.

Percentage of DIF Items

A high degree of agreement in DIF detection rates was observed between the 9% and 18% test at both levels of significance (see Table 16). More detailed results are shown in Tables 26 and 27 (Appendix A).

Table 16: Summary of percent detection rates for uniform DIF over 100 replications for the MH and LR procedures across percentages of DIF in the test.

% of DIF	$MH_{.05}$	$LR_{.05}$	$MH_{.01}$	$LR_{.01}$
9%	41.12	45.27	28.03	30.25
18%	36.86	40.77	21.72	24.42

Items in the 9% DIF test were identified marginally better than corresponding items in the 18% DIF test with both the MH and LR procedures. In 82% of the comparisons, items in the 9% DIF test were identified as well as or better than comparable items in the

18% DIF test. On average, items in the 9% test were identified 4-6% more frequently than in the 18% test. This difference in detection rates was highest with sample size 800, 5%, and decreased to 2% with sample size 100.

A higher percentage of DIF items would effect the ability scores of the focal group by making members of the focal group appear less able and, therefore, more likely not to respond correctly. Items may not appear to be biased because focal group members appear to be less able.

Rogers and Swaminathan (1993) found similar results. The detection rate of uniform DIF with the LR procedure increased by 6% when the percentage of DIF dropped from 15% to 0%, other than the item of interest. Rogers and Swaminathan (1993), however, found no effect due to percentage of DIF items on the detection rate with the MH procedure. Miller and Oshima (1992), on the other hand, found the DIF detection rate with the MH procedure to increase proportionately as the percentage of DIF items increased. According to the authors, these items had a moderate amount of DIF, b value difference was .35. DIF detection rates for items with small DIF, b value difference of .20, remained constant throughout changes in percentage of DIF items. Miller and Oshima (1992), however, used a wider range of percentage of DIF items, from 5% to 40% which may have had a bearing on their results.

False Positive Rates

False positive rates were determined by averaging the frequency with which non-DIF items were identified by either the MH or LR procedure. In the 9% DIF test, there were

60 non-DIF items; in the 18% DIF test, there were 54. In the following paragraphs, trends in the data related to false positives are discussed. False positive rates are summarized in Table 17. More detailed tables showing the false positive rates for the 9% and 18% DIF test at both levels of significance can be found in Appendix A (Tables 28, 29, 30, and 31).

Table 17: Mean percent false positive rates over 100 replications for the MH and LR procedures under all conditions.

Variable and Level	$MH_{.05}$	$LR_{.05}$	$MH_{.01}$	$LR_{.01}$
Sample Size				
100 Per Group	3.28	5.31	0.59	1.12
200 Per Group	4.28	5.78	0.86	1.24
400 Per Group	5.49	6.65	1.22	1.53
600 Per Group	6.19	7.32	1.41	1.80
800 Per Group	7.41	8.34	1.96	2.25
Percentage of DIF				
9%	4.44	5.66	0.90	1.23
18%	6.22	7.71	1.15	1.95
b value difference				
.2	4.10	5.19	0.74	1.01
.4	4.68	6.07	1.02	1.36
.6	5.61	7.03	1.27	1.70
.8	6.94	7.09	1.79	2.28

False positive rates for the MH and LR procedures were quite similar; however, the LR procedure consistently identified more false positives than the MH procedure. At the .05 level of significance, the LR procedure identified more than 1% more false positives than the MH procedure. Increasing sample size was accompanied by increasing false positive rates for both procedures.

The discrepancy in the false positive rates between the LR and MH procedures increased slightly as the sample size lowered. This relationship between the MH and LR false positive rates mirrored the relationship between the DIF detection rates. Tian, Pang, and Boss (1994b) also found that the LR false positive rate was slightly higher than the MH. At the .01 level of significance, the LR procedure identified approximately .4% more false positives than the MH procedure, averaged across sample size (Table 29, Appendix A). The discrepancy in the false positive rates between the LR and MH procedures also increased slightly as the sample size lowered.

As the amount of DIF (b value difference) increased in the items, the false positive rate for the other items on the test increased as well with both procedures over all conditions. This effect increased with sample size.

False positive rates were higher for the 18% DIF test than the 9% test for both the MH and LR procedure across all conditions. As the percentage of DIF items on the test increased, the scores of the reference group would tend to be inflated because of greater success on the DIF items. This would cause their scores to be more multidimensional. This effect would tend to be a problem when assessing DIF on non-DIF items. Therefore, these items might appear to be biased against the reference group. This effect was noticed, especially at larger sample sizes, when larger amounts of DIF were induced, and with a higher percentage of DIF items on the test.

CHAPTER V

SUMMARY AND CONCLUSIONS

In this chapter, the results are summarized by going back to the original research questions. The main purpose of the study was to examine the performance of the MH and LR procedures in the identification of uniform DIF in a number of different conditions. The detection rates for uniform DIF were quite similar throughout the manipulation of all variables; however, detection rates from the LR procedure were slightly higher than those from the MH procedure. The discrepancy between the detection rates decreased as sample size increased. False positive rates were also slightly higher with the LR procedure.

As the amount of DIF, measured in terms of b value difference and p-value difference, increased, the DIF detection rate of both procedures increased as well. However, effect size measured with p-value difference explained much more variance in MH and LR detection rates than did effect size measured with b value difference. P-value difference explained 95% of the variance in Δ_{MH}, compared to 60% from b value difference. Not only does p-value difference take into account b value difference but the a and b values of the item, as well. The results indicate that p-value difference would be a more accurate measure of effect size than b value difference. A p-value difference of .100 or larger was identified at least 90% of the time by both procedures at sample size 800 ($p<.05$). Averaged across all levels of sample size, a p-value difference of .100 or larger was identified at least 70% of the time by both procedures ($p<.05$) and approximately 60%

of the time at the .01 level. One of the recommendations from the study would be to report p-value difference when examining the influence of effect size on DIF detection rates when item parameters are known.

The effect of sample size on DIF detection was quite clear. DIF detection rates increased with larger sample sizes. Regression analysis demonstrated that sample size increased the effect of p-value difference. Furthermore, sample size had more of an effect on the MH detection rate than on the LR detection rate.

The question, 'How large a sample size is needed to detected different amounts of DIF?', depends on the accuracy wanted. The number of examinees used to detect DIF depends on the size of DIF that the test developer wishes to identify and the level of accuracy needed in the DIF detection rates. One would probably not want to use a sample size lower than 400. The regression models predicting MH and LR detection rates, presented in the study, may be of use when deciding on sample size for DIF identification purposes provided similar conditions are met.

Items with higher item discrimination were identified more frequently with both procedures. It would appear that this effect can be generalized for all values of a provided b is held constant. In the context of the study, higher levels of a combined with levels of b and b value difference to produce higher p-value differences. Conceivably, depending on the particular combination of a, b, and b value difference, an item with a larger a value may not be identified as frequently than one with a lower a value.

One must also be cautious to avoid over-simplifying the effect of item difficulty on DIF detection rates. In the present study, the item with the lower b value was detected

more frequently than the comparable item with a higher b value. This may not always be the case. Items with larger b values may have higher p-value differences than items with lower b values and, therefore, be detected more frequently. Again the effect of item difficulty on DIF detection rates depends on levels of a and b value difference and the direction of the DIF. One generalization about the effect of item difficulty on DIF detection rates was put forth regarding items of equal discrimination, equal DIF, and equal, but opposite, b values. If the DIF favours the reference group, the item with the negative b value will be detected more frequently because the p-value difference will be larger. If DIF favours the focal group, the item with the positive b value will be detected more frequently.

The percentage of DIF also had an effect on detection rates. Items on the 9% DIF test were identified marginally better than items on the 18% DIF test. DIF effects the ability scores of the focal group making the focal group appear less able. In turn, this may cause an item to appear unbiased when percentage of DIF items is large.

Limitations and Further Research

The greatest limitation of the study was the manipulation of the independent variables. In a simulation study all possible variables having an affect on the dependent variables are never truly taken into account. Only those variables that the author deems necessary are studied. As well, the levels of the variables chosen may not accurately reflect real data. In fact, the results obtained from the study can not really be generalized much further without additional examination. Further simulation studies examining the same variables, but at different levels, would be a beginning point with which to compare

the results of the study. Only after numerous simulated replications of the study would one want to generalize the results to real data. Additional variables such as unequal ability distribution or length of test could also yield interesting results.

REFERENCES

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

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.

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.

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

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).

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

Darlington, R.B. (1990). Regression and Linear Models. New York: McGraw-Hill Publishing Company.

Donoghue, H.R., & Allen, N.L. (1983). "Thin" versus "thick" matching in the Mantel-Haenszel procedure for detecting DIF. Journal of Educational Statistics, 18, 131-154.

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

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.

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

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).

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

Miller, M.D., & Oshima, T.C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. Applied Psychological Measurement, 16, 381-388.

Pang, X.L., Tian, F., & Boss, M.W. (1994). Performance of LR and MH DIF procedures over replications using real data. A paper presented at the Annual Meeting of American Educational Research Association. New Orleans.

Raju, N.S. (1988). The area between two item characteristic curves. Psychometrika, 53, 495-502.

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17, 105-116.

Spray, J.A. (1991). Estimation program of the logistic regression DIF statistic.

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

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

Tian, F., Pang, X.L., & Boss, M.W. (1994b). The effects of sample size and criterion variable on the identification of DIF by the Mantel-Haenszel and logistic regression procedures. A paper presented at the Annual Meeting of American Educational Research Association. New Orleans.

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

SUPPLEMENTARY TABLES

Table 18: Percent detection rates for uniform DIF over 100 replications for the MH and LR procedures for each p-value difference. Results are from the 9% DIF test ($p<.05$).

p diff. Ref. #	p diff.	100			200			400			600			800		
		Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR
1	.0167	-.07	1	0	-.12	4	5	-.15	6	6	-.19	8	12	-.18	15	15
2	.0202	-.32	3	11	-.16	2	4	-.25	15	18	-.20	9	9	-.19	9	11
3	.0294	-.26	3	3	-.24	3	6	-.31	12	13	-.30	18	19	-.31	21	22
4	.0300	-.30	3	3	-.37	10	14	-.27	4	9	-.34	25	25	-.34	26	26
5	.0351	-.35	3	6	-.48	10	15	-.41	17	20	-.41	29	30	-.39	31	33
6	.0365	-.28	3	6	-.29	8	11	-.38	17	18	-.37	22	24	-.29	23	27
7	.0396	-.35	6	11	-.38	9	9	-.48	19	20	-.39	23	28	-.43	36	38
8	.0405	-.40	8	10	-.34	7	11	-.42	20	24	-.38	22	25	-.37	28	35
9	.0560	-.46	6	7	-.52	15	21	-.54	34	36	-.52	37	43	-.56	56	62
10	.0579	-.37	8	13	-.57	11	21	-.53	31	32	-.55	43	52	-.52	51	53
11	.0604	-.74	10	18	-.63	15	21	-.69	38	44	-.66	58	60	-.57	60	64
12	.0609	-.54	9	16	-.58	19	24	-.57	36	39	-.57	48	51	-.58	61	61
13	.0687	-.77	11	17	-.81	27	30	-.74	51	55	-.71	64	67	-.79	89	92
14	.0751	-.66	10	18	-.76	22	27	-.60	36	47	-.65	59	62	-.69	78	82
15	.0767	-.85	12	22	-.81	22	29	-.84	56	65	-.86	75	80	-.87	86	90
16	.0812	-.79	11	21	-.82	33	39	-.77	57	59	-.78	73	79	-.74	79	83
17	.0855	-.81	10	17	-.78	24	38	-.81	55	61	-.84	80	85	-.81	91	93
18	.0942	-1.02	18	26	-.90	33	40	-.86	60	64	-.98	91	92	-.92	94	96
19	.1033	-1.09	26	34	-.98	42	50	-1.06	76	77	-1.04	88	90	-1.08	99	99
20	.1117	-.97	14	33	-1.04	44	56	-1.02	83	83	-1.07	96	97	-1.08	100	100
21	.1173	-1.34	25	39	-1.30	61	68	-1.26	88	94	-1.26	97	97	-1.34	100	100
22	.1246	-1.29	26	41	-1.17	55	68	-1.24	86	88	-1.19	100	100	-1.22	98	98
23	.1328	-1.47	37	52	-1.43	66	70	-1.39	98	98	-1.43	100	100	-1.44	100	100
24	.1544	-1.73	51	64	-1.70	85	91	-1.65	100	100	-1.68	100	100	-1.66	100	100

Where:

MH	LR U
1	0
51	64

Shaded areas are comparisons whereby the LR procedure identified the item less frequently than the MH procedure; representing 2% of the comparisons.

Unshaded areas are comparisons whereby the LR procedure identified the item as well as or better than the MH procedure; representing 98% of the comparisons.

Table 19: Percent detection rates for uniform DIF over 100 replications for the MH and LR procedures for each p-value difference. Results are from the 9% DIF test ($p<.01$).

p diff. Ref. #	p diff.	100			200			400			600			800		
		Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR
1	.0167	-.07	0	0	-.12	0	0	-.15	1	1	-.19	1	2	-.18	6	6
2	.0202	-.32	2	1	-.16	1	1	-.25	5	6	-.20	3	3	-.19	3	4
3	.0294	-.26	0	1	-.24	2	2	-.31	3	5	-.30	7	7	-.31	9	10
4	.0300	-.30	2	2	-.37	4	4	-.27	1	2	-.34	11	9	-.34	9	10
5	.0351	-.35	1	2	-.48	1	4	-.41	5	6	-.41	11	13	-.39	13	15
6	.0365	-.28	1	2	-.29	1	3	-.38	7	9	-.37	10	14	-.29	6	6
7	.0396	-.35	3	4	-.38	3	4	-.48	7	9	-.39	12	12	-.43	16	16
8	.0405	-.40	3	3	-.34	1	2	-.42	4	8	-.38	7	9	-.37	17	15
9	.0560	-.46	2	5	-.52	3	5	-.54	17	16	-.52	18	18	-.56	35	38
10	.0579	-.37	1	4	-.57	5	6	-.53	12	13	-.55	20	25	-.52	39	39
11	.0604	-.74	6	8	-.63	6	9	-.69	19	25	-.66	31	37	-.57	37	38
12	.0609	-.54	2	2	-.58	4	7	-.57	10	14	-.57	24	27	-.58	40	42
13	.0687	-.77	3	7	-.81	14	14	-.74	27	29	-.71	40	41	-.79	64	66
14	.0751	-.66	2	3	-.76	12	17	-.60	17	24	-.65	39	47	-.69	61	64
15	.0767	-.85	1	6	-.81	10	13	-.84	24	32	-.86	53	56	-.87	72	75
16	.0812	-.79	3	6	-.82	13	16	-.77	31	38	-.78	50	55	-.74	60	62
17	.0855	-.81	3	8	-.78	11	14	-.81	34	37	-.84	57	56	-.81	71	73
18	.0942	-1.02	9	11	-.90	16	19	-.86	38	42	-.98	66	73	-.92	81	83
19	.1033	-1.09	11	17	-.98	18	28	-1.06	59	62	-1.04	74	77	-1.08	96	95
20	.1117	-.97	5	10	-1.04	24	30	-1.02	50	61	-1.07	86	90	-1.08	95	97
21	.1173	-1.34	10	17	-1.30	39	42	-1.26	71	81	-1.26	93	95	-1.34	98	99
22	.1246	-1.29	11	18	-1.17	34	44	-1.24	76	82	-1.19	91	92	-1.22	97	97
23	.1328	-1.47	11	26	-1.43	43	54	-1.39	84	87	-1.43	99	100	-1.44	100	100
24	.1544	-1.73	22	34	-1.70	62	74	-1.65	93	97	-1.68	100	100	-1.66	100	100

Where:

MH	LR U
2	1
22	34

Shaded areas are comparisons whereby the LR procedure identified the item less frequently than the MH procedure; representing 5% of the comparisons.

Unshaded areas are comparisons whereby the LR procedure identified the item as well as or better than the MH procedure; representing 95% of the comparisons.

Table 20: Percent detection rates for uniform DIF over 100 replications for the MH and LR procedures for each p-value difference. Results are from the 18% DIF test ($p<.05$).

Table continues on next page.

p diff. Ref. #	p diff.	100			200			400			600			800		
		Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR
1	.0167	-.16	3	5	-.21	6	7	-.16	7	9	-.17	10	13	-.16	9	9
		-.31	4	7	-.24	5	5	-.14	8	11	-.18	9	8	-.13	5	5
2	.0202	.00	0	4	-.14	3	4	-.19	5	7	-.16	10	10	-.17	7	7
		-.14	5	5	-.17	6	10	-.19	4	8	-.12	6	6	-.20	11	14
3	.0294	-.27	2	3	-.35	5	10	-.27	12	15	-.24	9	8	-.27	21	25
		-.35	7	8	-.29	6	9	-.25	6	11	-.31	14	17	-.29	23	22
4	.0300	-.34	1	13	-.27	5	6	-.28	13	17	-.24	9	11	-.27	18	19
		-.26	7	7	-.30	7	8	-.27	12	9	-.32	16	20	-.32	24	25
5	.0351	-.21	2	4	-.37	15	20	-.37	16	18	-.35	16	18	-.37	25	27
		-.42	4	6	-.29	3	4	-.35	16	18	-.33	17	19	-.37	26	31
6	.0365	-.18	4	5	-.25	8	10	-.37	15	15	-.32	20	27	-.34	25	28
		-.32	5	10	-.28	8	12	-.35	12	16	-.31	13	14	-.32	25	28
7	.0396	-.36	3	4	-.44	9	13	-.39	14	16	-.43	20	21	-.42	30	33
		-.40	2	11	-.38	7	11	-.40	13	18	-.40	19	22	-.36	24	26
8	.0405	-.39	3	7	-.37	10	16	-.34	13	14	-.34	17	22	-.31	22	23
		-.37	3	8	-.37	8	11	-.39	16	23	-.34	27	25	-.33	28	29
9	.0560	-.39	4	7	-.47	8	13	-.45	22	22	-.48	34	38	-.48	51	51
		-.45	12	14	-.34	9	11	-.43	22	27	-.52	42	47	-.46	52	54
10	.0579	-.36	4	5	-.50	14	20	-.53	26	32	-.51	38	44	-.51	52	56
		-.75	12	15	-.40	11	15	-.48	25	27	-.54	40	42	-.50	48	50
11	.0604	-.57	12	13	-.49	9	13	-.53	31	33	-.57	40	44	-.59	57	60
		-.57	6	13	-.50	13	18	-.58	34	39	-.56	43	46	-.58	51	54
12	.0609	-.48	8	11	-.53	13	20	-.51	25	33	-.53	38	44	-.50	51	58
		-.57	10	12	-.55	14	15	-.51	26	28	-.55	42	49	-.54	53	59

P diff. Ref. #	P diff.	100			200			400			600			800		
		Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR
13	.0687	-.77	14	18	-.66	17	20	-.63	43	44	-.63	55	58	-.63	65	70
		-.66	13	18	-.71	20	25	-.64	31	36	-.61	45	53	-.66	71	71
14	.0751	-.59	13	18	-.63	14	25	-.62	37	48	-.60	57	60	-.64	71	77
		-.54	5	10	-.64	21	29	-.61	37	41	-.59	49	56	-.64	70	73
15	.0767	-.77	7	17	-.81	26	37	-.75	54	57	-.78	66	73	-.78	83	84
		-.68	12	16	-.83	25	32	-.81	51	54	-.80	70	77	-.73	74	76
16	.0812	-.76	15	24	-.74	31	34	-.70	43	50	-.63	58	63	-.68	78	80
		-.63	10	12	-.63	18	21	-.69	46	46	-.68	66	69	-.65	69	74
17	.0855	-.78	11	18	-.67	22	30	-.72	53	56	-.69	64	67	-.71	80	83
		-.60	10	12	-.68	21	25	-.78	48	57	-.76	73	74	-.70	79	80
18	.0942	-.75	14	21	-.78	27	34	-.81	56	63	-.83	75	77	-.83	83	88
		-.92	23	25	-.82	28	37	-.89	66	64	-.82	77	94	-.84	91	91
19	.1033	-.99	21	26	-.90	40	43	-1.02	76	78	-.89	81	85	-.95	92	93
		-.98	9	25	-.94	34	45	-.98	75	76	-.97	84	86	-1.00	96	97
20	.1117	-1.05	24	28	-.94	35	40	-.94	72	75	-.92	88	90	-.92	97	97
		-.92	13	24	-.94	39	49	-.89	65	73	-.98	89	91	-.94	94	95
21	.1173	-1.09	17	32	-1.14	44	55	-1.18	83	88	-1.10	97	98	-1.14	98	99
		-1.26	24	39	-1.13	43	58	-1.15	86	90	-1.21	98	98	-1.14	97	99
22	.1246	-1.11	23	37	-1.09	50	56	-1.10	78	85	-1.13	96	97	-1.11	96	96
		-1.15	22	38	-1.17	50	61	-1.10	82	87	-1.12	96	97	-1.12	99	100
23	.1328	-1.12	22	32	-1.15	50	62	-1.20	84	90	-1.22	96	99	-1.25	100	100
		-1.31	24	40	-1.26	64	70	-1.22	88	91	-1.23	100	100	-1.25	100	100
24	.1544	-1.53	35	45	-1.59	79	84	-1.56	99	99	-1.48	100	100	-1.53	100	100
		-1.53	35	50	-1.58	74	83	-1.48	97	99	-1.56	100	100	-1.50	100	100

Where:

MH	LR U
0	0
35	50

Shaded areas are comparisons whereby the LR procedure identified the item less frequently than the MH procedure; representing 3% of the comparisons.

Unshaded areas are comparisons whereby the LR procedure identified the item as well as or better than the MH procedure; representing 97% of the comparisons.

Table 21: Percent detection rates for uniform DIF over 100 replications for the MH and LR procedures for each p-value difference. Results are from the 18% DIF test (p<.01).

Table continues on next page.

p diff. Ref. #	p diff.	100			200			400			600			800		
		Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR
1	.0167	-.16	1	2	-.21	0	1	-.16	1	2	-.17	4	6	-.16	1	2
		-.31	0	3	-.24	1	1	-.14	0	1	-.18	2	2	-.13	2	1
2	.0202	.00	0	2	-.14	1	2	-.19	0	0	-.16	4	4	-.17	3	4
		-.14	0	1	-.17	1	2	-.19	1	1	-.12	2	3	-.20	3	5
3	.0294	-.27	0	1	-.35	2	2	-.27	2	3	-.24	2	4	-.27	4	7
		-.35	3	4	-.29	0	3	-.25	1	1	-.31	4	4	-.29	7	6
4	.0300	-.34	1	6	-.27	2	2	-.28	5	8	-.24	6	7	-.27	5	6
		-.26	5	2	-.30	2	2	-.27	1	1	-.32	6	6	-.32	10	10
5	.0351	-.21	1	2	-.37	6	5	-.37	8	10	-.35	6	7	-.37	10	13
		-.42	0	2	-.29	1	2	-.35	6	10	-.33	7	7	-.37	12	12
6	.0365	-.18	1	2	-.25	1	3	-.37	3	3	-.32	11	11	-.34	11	11
		-.32	3	2	-.28	2	3	-.35	1	3	-.31	7	7	-.32	12	12
7	.0396	-.36	0	0	-.44	1	4	-.39	4	5	-.43	10	10	-.42	9	11
		-.40	0	1	-.38	1	3	-.40	5	3	-.40	5	4	-.36	9	10
8	.0405	-.39	1	3	-.37	1	2	-.34	3	4	-.34	8	9	-.31	9	13
		-.37	1	1	-.37	3	4	-.39	5	7	-.34	8	12	-.33	10	13
9	.0560	-.39	0	1	-.47	0	1	-.45	12	15	-.48	20	22	-.48	25	24
		-.45	2	2	-.34	2	5	-.43	7	11	-.52	24	26	-.46	26	28
10	.0579	-.36	3	2	-.50	5	4	-.53	9	12	-.51	26	27	-.51	25	31
		-.75	6	8	-.40	4	3	-.46	8	10	-.54	19	21	-.50	26	31
11	.0604	-.57	3	6	-.49	3	4	-.53	10	14	-.57	22	26	-.59	37	40
		-.57	3	3	-.50	5	8	-.58	9	18	-.56	20	17	-.58	33	38
12	.0609	-.48	2	4	-.53	5	6	-.51	16	15	-.53	17	17	-.50	21	25
		-.57	1	1	-.55	8	7	-.51	8	11	-.55	23	27	-.54	27	29

p diff. Ref. #	p diff.	100			200			400			600			800		
		Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR
13	.0687	-.77	7	7	-.66	4	5	-.63	19	21	-.63	29	28	-.63	40	43
		-.66	3	4	-.71	11	12	-.64	17	21	-.61	23	26	-.66	42	49
14	.0751	-.59	4	10	-.63	8	9	-.62	24	28	-.60	30	36	-.64	50	51
		-.54	0	2	-.64	7	8	-.61	16	19	-.59	32	33	-.64	53	52
15	.0767	-.77	1	6	-.81	7	9	-.75	25	28	-.78	48	49	-.78	56	62
		-.68	3	7	-.83	7	15	-.81	22	27	-.80	44	54	-.73	54	55
16	.0812	-.76	6	11	-.74	11	17	-.70	23	28	-.63	36	39	-.68	53	63
		-.63	2	5	-.63	11	14	-.69	29	30	-.68	38	46	-.65	48	50
17	.0855	-.78	5	6	-.67	8	12	-.72	29	34	-.69	38	44	-.71	58	61
		-.60	3	7	-.68	12	13	-.78	36	36	-.76	52	55	-.70	58	58
18	.0942	-.75	2	6	-.78	14	22	-.81	25	37	-.83	57	59	-.83	66	70
		-.92	7	9	-.82	11	19	-.89	35	44	-.82	51	55	-.84	77	80
19	.1033	-.99	11	16	-.90	21	25	-1.02	51	54	-.89	64	71	-.95	82	84
		-.98	2	8	-.94	13	18	-.98	48	57	-.97	73	76	-1.00	85	85
20	.1117	-1.05	6	14	-.94	18	24	-.94	44	55	-.92	74	76	-.92	87	88
		-.92	4	8	-.94	17	24	-.89	44	52	-.98	78	80	-.94	83	84
21	.1173	-1.09	7	11	-1.14	23	29	-1.18	64	69	-1.10	82	88	-1.14	93	96
		-1.26	7	19	-1.13	24	31	-1.15	62	74	-1.21	90	93	-1.14	95	95
22	.1246	-1.11	9	13	-1.09	19	28	-1.10	60	63	-1.13	90	91	-1.11	92	93
		-1.15	7	15	-1.17	30	35	-1.10	59	72	-1.12	82	86	-1.12	98	98
23	.1328	-1.12	6	14	-1.15	27	39	-1.20	68	72	-1.22	94	95	-1.25	99	99
		-1.31	11	17	-1.26	38	47	-1.22	76	82	-1.23	94	94	-1.25	98	98
24	.1544	-1.53	17	29	-1.59	54	63	-1.56	93	97	-1.48	98	99	-1.53	100	100
		-1.53	13	30	-1.58	57	66	-1.48	89	92	-1.56	100	100	-1.50	100	100

Where:

MH	LR U
2	1
13	30

Shaded areas are comparisons whereby the LR procedure identified the item less frequently than the MH procedure; representing 5% of the comparisons.

Unshaded areas are comparisons whereby the LR procedure identified the item as well as or better than the MH procedure; representing 95% of the comparisons.

Table 22: Percent detection rates for uniform DIF over 100 replications for the MH and LR procedures averaged across levels of b value difference, a value, and b value. Results are from the 9% DIF test (p<.05).

Variable	Level	100			200			400			600			800		
		Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR
b value difference	.2	-0.27	3.17	5.67	-0.29	6.33	8.83	-0.31	12.17	14.33	-0.30	18.67	20.50	-0.31	23.00	24.17
	.4	-0.57	8.64	14.33	-0.58	15.00	20.50	-0.60	35.50	39.67	-0.59	47.33	51.33	-0.57	56.17	60.17
	.6	-0.88	15.67	23.17	-0.84	32.33	40.17	-0.85	58.17	61.83	-0.87	73.50	76.33	-0.88	83.50	85.17
	.8	-1.15	24.83	38.17	-1.15	50.83	58.50	-1.11	76.67	79.17	-1.13	88.33	89.67	-1.14	92.50	93.83
a value	.3	-0.44	6.38	11.13	-0.45	13.75	17.75	-0.46	27.63	30.88	-0.46	34.75	38.13	-0.45	43.63	47.00
	.5	-0.72	11.50	19.25	-0.71	24.38	33.00	-0.72	46.13	49.25	-0.74	64.13	66.25	-0.72	67.63	69.00
	.7	-0.99	21.38	30.63	-0.99	40.25	45.25	-0.98	63.13	66.13	-0.97	72.00	74.00	-1.00	80.13	81.50
b value RG	-.5	-0.81	15.17	23.50	-0.76	29.25	34.83	-0.78	48.25	52.00	-0.77	60.08	62.17	-0.77	64.75	66.83
	.5	-0.63	11.00	17.17	-0.67	23.00	29.17	-0.66	43.00	45.50	-0.67	53.38	56.75	-0.68	62.83	64.83

Table 23: Percent detection rates for uniform DIF over 100 replications for the MH and LR procedures averaged across levels of b value difference, a value, and b value. Results are from the 9% DIF test (p<.01).

Variable	Level	100			200			400			600			800		
		Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR
b value difference	.2	-0.27	1.33	1.67	-0.29	1.83	2.50	-0.31	3.67	4.83	-0.30	7.50	7.67	-0.31	9.33	10.17
	.4	-0.57	2.50	5.00	-0.58	6.17	7.83	-0.60	15.50	19.33	-0.59	26.83	30.33	-0.57	39.17	39.67
	.6	-0.88	6.17	10.00	-0.84	15.17	19.17	-0.85	38.17	42.00	-0.87	55.33	57.67	-0.88	70.17	71.67
	.8	-1.15	9.00	16.17	-1.15	31.33	39.17	-1.11	58.50	64.83	-1.13	77.50	80.67	-1.14	85.50	86.67
a value	.3	-0.44	1.88	2.75	-0.45	4.38	6.38	-0.46	11.50	14.50	-0.46	19.00	21.88	-0.45	28.50	29.50
	.5	-0.72	4.63	7.75	-0.71	12.75	16.00	-0.72	29.13	33.38	-0.74	46.13	48.63	-0.72	54.75	55.88
	.7	-0.99	7.75	14.13	-0.99	23.75	29.13	-0.98	46.25	50.38	-0.97	60.25	61.75	-1.00	69.88	70.75
b value RG	-.5	-0.81	6.17	9.33	-0.76	16.08	19.58	-0.78	31.58	36.33	-0.77	45.08	47.33	-0.77	52.50	53.42
	.5	-0.63	3.33	7.08	-0.67	11.17	14.75	-0.66	26.33	29.17	-0.37	38.50	40.83	-0.68	49.58	50.67

Table 24: Percent detection rates for uniform DIF over 100 replications for the MH and LR procedures averaged across levels of b value difference, a value, and b value. Results are from the 18% DIF test (p<.05).

Variable	Level	100			200			400			600			800		
		Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR
b value difference	.2	-0.27	3.33	6.42	-0.29	6.42	8.92	-0.27	10.50	13.08	-0.27	12.92	14.42	-0.28	18.50	20.25
	.4	-0.53	7.92	12.08	-0.51	14.08	19.08	-0.53	29.25	32.50	-0.53	39.50	43.75	-0.52	50.08	52.42
	.6	-0.77	13.58	20.17	-0.75	25.25	32.17	-0.79	53.17	56.83	-0.38	67.08	71.42	-0.77	77.08	79.33
	.8	-1.02	20.08	29.83	-1.03	43.75	51.17	-1.01	69.00	73.67	-1.01	82.92	85.17	-1.02	89.50	91.00
a value	.3	-0.39	6.50	9.94	-0.41	11.38	15.19	-0.42	21.13	24.88	-0.41	31.13	34.44	-0.41	39.19	41.81
	.5	-0.67	11.94	17.50	-0.64	21.38	26.94	-0.65	42.44	46.44	-0.66	54.19	57.44	-0.66	63.31	65.06
	.7	-0.88	15.25	23.94	-0.89	34.38	41.38	-0.88	57.88	60.75	-0.87	66.50	69.19	-0.88	73.94	75.38
b value RG	-.5	-0.69	12.38	19.33	-0.70	24.96	30.71	-0.70	43.63	46.96	-0.70	53.58	56.79	-0.69	60.17	62.25
	.5	-0.60	10.08	14.92	-0.59	19.79	24.96	-0.60	37.33	41.08	-0.60	47.63	50.58	-0.60	57.46	59.25

Table 25: Percent detection rates for uniform DIF over 100 replications for the MH and LR procedures averaged across levels of b value difference, a value, and b value. Results are from the 18% DIF test (p<.01).

Variable	Level	100			200			400			600			800		
		Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR	Δ_{MH}	MH	LR
b value difference	.2	-0.27	0.92	2.17	-0.29	1.50	2.42	-0.27	2.83	3.75	-0.27	4.83	5.33	-0.28	6.25	7.25
	.4	-0.53	2.92	4.25	-0.51	4.42	6.00	-0.53	10.92	14.00	-0.53	22.08	23.92	-0.52	29.58	33.17
	.6	-0.77	4.08	7.50	-0.75	11.75	15.67	-0.79	32.75	38.08	-0.78	49.25	52.75	-0.77	59.42	61.25
	.8	-1.02	7.08	14.00	-1.03	24.75	31.17	-1.01	52.08	57.50	-1.01	70.50	72.92	-1.02	80.08	81.33
a value	.3	-0.39	1.50	3.25	-0.41	3.88	5.31	-0.42	9.31	11.13	-0.41	16.63	18.75	-0.41	22.13	23.94
	.5	-0.67	4.19	6.88	-0.64	9.50	12.81	-0.65	23.56	28.75	-0.66	39.19	41.13	-0.66	47.88	50.06
	.7	-0.88	5.56	10.81	-0.89	18.44	23.31	-0.88	41.06	45.13	-0.87	54.19	56.31	-0.88	61.50	63.25
b value RG	-.5	-0.69	4.08	7.96	-0.70	12.54	16.42	-0.70	27.21	31.17	-0.70	39.46	41.71	-0.69	46.17	48.58
	.5	-0.60	3.42	6.00	-0.59	8.67	11.21	-0.60	22.08	25.50	-0.60	33.88	35.75	-0.60	41.50	42.92

Table 26: Percent detection rates for uniform DIF by percent of DIF items for sample size and p-value difference over 100 replications for the MH and LR procedures (p<.05).

Table continues on next page.

p diff. Ref. #	p diff.	% DIF	100		200		400		600		800	
			MH	LR	MH	LR	MH	LR	MH	LR	MH	LR
1	.0167	9%	1	0	4	5	6	6	8	12	15	15
		18%	3.5	6	5.5	6	7.5	10	9.5	10.5	7	7
2	.0202	9%	3	11	2	4	15	18	9	9	9	11
		18%	2.5	4.5	4.5	7	4.5	7.5	8	8	9	10.5
3	.0294	9%	3	3	3	6	12	13	18	19	21	22
		18%	4.5	5.5	5.5	9.5	9	13	11.5	12.5	21.5	23.5
4	.0300	9%	3	3	10	14	4	9	25	25	26	26
		18%	4	10.5	6	7	12.5	13	12.5	15.5	21	22
5	.0351	9%	3	6	10	15	17	20	29	30	31	33
		18%	3	5	9	12	16	18	16.5	18.5	25.5	29
6	.0365	9%	3	6	8	11	17	18	22	24	23	27
		18%	4.5	7.5	8	11	13.5	15.5	16.5	20.5	25	28
7	.0396	9%	6	11	9	9	19	20	23	28	36	38
		18%	2.5	7.5	8	12	13.5	17	19.5	21.5	37	29.5
8	.0405	9%	8	10	7	11	20	24	22	25	28	35
		18%	3	7.5	9	13.5	14.5	18.5	22	23.5	25	26
9	.0560	9%	6	7	15	21	34	36	37	43	56	62
		18%	6	10.5	8.5	12	22	24.5	36	42.5	51.5	52.5
10	.0579	9%	8	13	11	21	31	32	43	52	51	53
		18%	8	10	12.5	17.5	25.5	29.5	39	43	50	53
11	.0604	9%	10	18	15	21	38	44	58	60	60	64
		18%	9	13	11	15.5	32.5	36	41.5	45	54	57
12	.0609	9%	9	16	19	24	36	39	48	51	61	61
		18%	9	11.5	13.5	17.5	25.5	30.5	40	46.5	52	58.5

p diff. Ref. #	p diff.	% DIF	100		200		400		600		800	
			MH	LR	MH	LR	MH	LR	MH	LR	MH	LR
13	.0687	9%	11	17	27	30	51	55	64	67	89	92
		18%	13.5	18	18.5	22.5	47	40	50	55.5	68	70.5
14	.0751	9%	10	18	22	27	36	47	59	62	78	82
		18%	9	14	17.5	27	37	44.5	53	58	70.5	75
15	.0767	9%	12	22	22	29	56	65	75	80	86	90
		18%	9.5	16.5	25.5	34.5	52.5	55.5	68	75	78.5	80
16	.0812	9%	11	21	33	39	57	59	73	79	79	83
		18%	12.5	18	24.5	27.5	44.5	48	62	66	73.5	87
17	.0855	9%	10	17	24	38	55	61	80	85	91	93
		18%	10.5	15	21.5	27.5	50.5	56.5	68.5	70.5	79.5	81.5
18	.0942	9%	18	26	33	40	60	64	91	92	94	96
		18%	18.5	23	27.5	35.5	61	63.5	76	85.5	87	89.5
19	.1033	9%	26	34	42	50	76	77	88	90	99	99
		18%	15	25.5	37	44	75.5	77	82.5	85.5	95	95
20	.1117	9%	14	33	44	56	83	83	98	97	100	100
		18%	18.5	26	37	44.5	68.5	74	88.5	90.5	95.5	98
21	.1173	9%	25	39	61	68	88	94	97	97	100	100
		18%	20.5	35.5	43.5	56.5	84.5	89	97.5	98	97.5	99
22	.1246	9%	26	41	55	68	86	88	100	100	98	98
		18%	22.5	37.5	50	58.5	80	86	96	97	97.5	98
23	.1328	9%	37	52	66	70	98	98	100	100	100	100
		18%	23	36	57	66	86	90.5	98	99.5	100	100
24	.1544	9%	51	64	85	91	100	100	100	100	100	100
		18%	35	47.5	76.5	83.5	98	99	100	100	100	100

Where:

9%	1
18%	3.5
9%	51
18%	35

Shaded areas are comparisons whereby the item in the 9% DIF test was identified less frequently than the comparable item in the 18% DIF test; representing 18% of the comparisons.

Unshaded areas are comparisons whereby the item in the 9% DIF test was identified as well as or better than the comparable item in the 18% DIF test; representing 82% of the comparisons.

Totals for the 18% DIF test are averages of the two items at the same p-value difference.

Table 27: Percent detection rates for uniform DIF by percent of DIF items for sample size and p-value difference over 100 replications for the MH and LR procedures (p<.01).

Table continues on next page.

p diff. Ref. #	p diff.	% DIF	100		200		400		600		800	
			MH	LR	MH	LR	MH	LR	MH	LR	MH	LR
1	.0167	9%	0	0	0	0	1	1	1	2	6	6
		18%	.5	2.5	.5	1	.5	1.5	3	4	1.5	1.5
2	.0202	9%	2	1	1	1	5	6	3	3	3	4
		18%	0	1.5	1	2	.5	.5	3	3.5	3	4.5
3	.0294	9%	0	1	2	2	3	5	7	7	9	10
		18%	1.5	2.5	1	2.5	1.5	2	3	4	5.5	6.5
4	.0300	9%	2	2	4	4	1	2	11	9	9	10
		18%	3	4	2	2	3	4.5	6	6.5	7.5	8
5	.0351	9%	1	2	1	4	5	6	11	13	13	15
		18%	.5	2	3.5	3.5	7	10	6.5	7	11	12.5
6	.0365	9%	1	2	1	3	7	9	10	14	6	5
		18%	2	2	1.5	3	2	3	9	9	11.5	11.5
7	.0396	9%	3	4	3	4	7	9	12	12	16	16
		18%	0	.5	1	3.5	4.5	4	7.5	7	9	10.5
8	.0405	9%	3	3	1	2	4	8	7	9	17	15
		18%	1	2	2	3	4	5.5	6	10.5	9.5	13
9	.0560	9%	2	5	3	5	17	16	18	18	35	38
		18%	1	1.5	1	3	9.5	13	22	24	25.5	26
10	.0579	9%	1	4	5	6	12	13	20	25	39	39
		18%	4.5	6	4.5	3.5	6.5	11	22.5	24	25.5	31
11	.0604	9%	6	8	6	9	19	25	31	37	37	38
		18%	3	4.5	4	6	9.5	16	21	21.5	35	39
12	.0609	9%	2	2	4	7	10	14	24	27	40	42
		18%	1.5	2.5	6.5	6.5	12	13	20	22	24	27

p diff. Ref. #	p diff.	% DIF	100		200		400		600		800	
			MH	LR	MH	LR	MH	LR	MH	LR	MH	LR
13	.0687	9%	3	7	14	14	27	29	40	41	64	66
		18%	5	5.5	7.5	8.5	18	21	26	27	41	46
14	.0751	9%	2	3	12	17	17	24	39	47	61	64
		18%	2	6	7.5	8.5	20	23.5	31	34.5	51.5	51.5
15	.0767	9%	1	6	10	13	24	32	53	56	72	75
		18%	2	6.5	7	12	23.5	27.5	46	51.5	55	58.5
16	.0812	9%	3	6	13	16	31	38	50	55	60	62
		18%	4	8	11	15.5	26	29	37	32.5	50.5	56.5
17	.0855	9%	3	8	11	14	34	37	57	56	71	73
		18%	4	6.5	10	12.5	22.5	35	45	49.5	58	59.5
18	.0942	9%	9	11	16	19	38	42	66	73	81	83
		18%	4.5	7	12.5	20.5	30	40.5	54	57	71.5	75
19	.1033	9%	11	17	18	28	59	62	74	77	96	95
		18%	6.5	12	17	21.5	49.5	55.5	68.5	73.5	83.5	84.5
20	.1117	9%	5	10	24	30	50	61	86	90	95	97
		18%	5	11	17.5	24	44	53.5	76	78	85	86
21	.1173	9%	10	17	39	42	71	81	93	95	98	99
		18%	7	15	23.5	30	63	71.5	86	90.5	94	95.5
22	.1246	9%	11	18	34	44	76	82	91	92	97	97
		18%	8	14	24.5	31.5	59.5	67.5	86	88.5	95	95.5
23	.1328	9%	11	26	43	54	84	87	99	100	100	100
		18%	8.5	15.5	32.5	43	72	77	94	94.5	98.5	98.5
24	.1544	9%	22	34	62	74	93	97	100	100	100	100
		18%	15	29.5	55.5	64.5	91	94.5	99	99.5	100	100

Where:

9%	3
18%	5
9%	22
18%	15

Shaded areas are comparisons whereby the item in the 9% DIF test was identified less frequently than the comparable item in the 18% DIF test; representing 19% of the comparisons.

Unshaded areas are comparisons whereby the item in the 9% DIF test was identified as well as or better than the comparable item in the 18% DIF test; representing 81% of the comparisons.

Totals for the 18% DIF test are averages of the two items at the same p-value difference.

Table 28: Percent false positive rates over 100 replications for the MH and LR procedures. Results are from the 9% DIF test (p<.05).

b-diff	100		200		400		600		800	
	MH	LR	MH	LR	MH	LR	MH	LR	MH	LR
.2	3.05	4.77	3.48	4.82	4.25	5.02	4.07	4.67	4.52	5.28
.4	3.05	5.17	4.45	5.72	4.18	5.22	4.33	5.47	4.62	5.35
.6	3.50	5.38	3.53	5.10	4.88	5.77	4.90	6.18	5.97	6.85
.8	3.15	5.55	4.25	5.70	5.35	5.93	6.12	7.10	7.20	8.08
MEAN	3.19	5.22	3.93	5.33	4.67	5.48	4.85	5.85	5.58	6.39

Table 29: Percent false positive rates over 100 replications for the MH and LR procedures. Results are from the 9% DIF test (p<.01).

b-diff	100		200		400		600		800	
	MH	LR	MH	LR	MH	LR	MH	LR	MH	LR
.2	0.53	0.93	0.52	0.80	0.68	0.90	0.67	0.95	0.98	1.22
.4	0.60	1.08	0.87	1.50	0.90	0.98	0.77	1.10	1.03	1.25
.6	0.65	1.28	0.60	0.88	1.07	1.33	1.17	1.38	1.32	1.57
.8	0.58	1.05	0.70	1.12	1.23	1.50	1.23	1.58	1.92	2.12
MEAN	0.59	1.09	0.67	1.07	0.97	1.18	0.96	1.25	1.31	1.54

Table 30: Percent false positive rates over 100 replications for the MH and LR procedures. Results are from the 18% DIF test (p<.05).

b-diff	100		200		400		600		800	
	MH	LR	MH	LR	MH	LR	MH	LR	MH	LR
.2	2.98	4.52	4.28	5.61	4.61	5.89	4.43	5.04	5.32	6.30
.4	3.52	5.74	3.74	5.45	5.83	7.22	5.94	7.54	7.17	7.80
.6	3.48	5.94	4.59	6.11	6.55	8.22	8.52	9.59	10.13	11.13
.8	3.54	5.37	5.98	7.74	8.28	9.96	11.20	13.00	14.35	15.96
MEAN	3.38	5.39	4.65	6.23	6.32	7.82	7.52	8.79	9.24	10.30

Table 31: Percent false positive rates over 100 replications for the MH and LR procedures. Results are from the 18% DIF test (p<.01).

b-diff	100		200		400		600		800	
	MH	LR	MH	LR	MH	LR	MH	LR	MH	LR
.2	0.48	0.83	0.81	0.98	0.74	1.02	0.81	1.11	1.20	1.31
.4	0.65	1.13	0.85	1.02	1.48	1.91	1.31	1.67	1.78	1.98
.6	0.52	1.24	1.13	1.69	1.56	1.93	2.07	2.57	2.61	3.09
.8	0.74	1.370.	1.37	1.96	2.09	2.65	3.22	4.04	4.83	5.45
MEAN	0.60	1.14	1.04	1.41	1.47	1.87	1.86	2.35	2.61	2.96