The Performance of the Mantel-Haenszel and Logistic Regression Dif Detection Procedures Across Sample Size and Effect Size: A Monte Carlo Study

Patrick Hadley

Thesis submitted to
the School of Graduate Studies and Research
in partial fulfillment of the requirements fo the M.A.
degree in Education

University of Ottawa

1994

ISBN 0-612-04875-6

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ABSTRACT

In recent years, public attention has become focused on the issue of test and item bias in standardized tests. Since the 1980's, the Mantel-Haenszel (Holland & Thayer, 1986) and Logistic Regression procedures (Swaminathan & Rogers, 1990) have been developed to detect item bias, or differential item functioning (dif). In this study the effectiveness of the MH and LR procedures was compared under a variety of conditions, using simulated data.

The ability of the MH and LR to detect dif was tested at sample sizes of 100/100, 200/200, 400/400, 600/600, and 800/800. The simulated test had 66 items, the first 33 items with item discrimination ("a") set at 0.80, the second 33 items with "a" set at 1.20. The pseudo-guessing parameter ("c") was 0.15 for all items. The item difficulty ("b") parameter ranged from -2.00 to 2.00 in increments of 0.125 for the first 33 items, and again for the second 33 items. Four dif items, two from each half of the test, with "b" levels of -0.75 and 0.75 for the reference group, were set so that "b" was .25, .5., .75, and 1.00 higher for examinees in the focal group. This difference was translated into differences in p-values, and the p-values were arcsine transformed into "h", using the method proposed by Cohen (1992), to provide two other ways to measure degree of dif, or "effect size". One hundred replications were run, and detection and false-positive rates were tabulated for both procedures. Regression models were constructed so as to predict MH and LRU dif detection rates, as well as the mean value of Δ_{MH} (the Mantel-Haenszel delta) across replications.

Both the MH and LRU detected dif with a high degree of success whenever sample size was large (600 or more), especially when effect size, no matter how measured, was also large.

Only when sample size was less than 600 did either the MH or LRU have difficulty detecting dif items when effect size was large. The MH and LRU were more able to detect items of low difficulty than high difficulty, and of high discrimination than low discrimination. The LRU outperformed the MH marginally under almost every condition of the study. However, the LRU also had a higher false-positive rate than the MH, a finding consistent with previous studies (Pang et al., 1994, Tian et al., 1994a, 1994b). It was concluded that the MH and LRU were generally reliable detectors of dif at sample sizes of about 500 or larger.

The Δ_{MH} was found to be so highly correlated with "h", that "h" was sufficient to provide a complete regression model for the prediction of mean Δ_{MH} value across replications. The conclusion was drawn that the Δ_{MH} is itself an excellent measure of effect size. Since the "a" and "b" parameters which underly the computation of the three measures of effect size used in the study are not always determinable in data derived from real world test administrations, it may be that the Δ_{MH} is the best available measure of effect size in real world test items.

Acknowledgements

I wish to thank Dr. Marvin Boss, my thesis supervisor, for his invaluable guidance throughout the entire process of writing this thesis. His suggestions, criticisms, and encouragement were all of inestimable value, but it was his expertise that was perhaps the single most important help to me over the last year. Writing a thesis is made much easier when you have an adviser of the calibre of Dr. Boss.

Drs. Marc Gessaroli and Bruno Zumbo offered extremely useful feedback at several critical points, and their advice was greatly appreciated.

To Chris Carruthers for his much-needed assistance in the debugging of a very complex computer program, my heartfelt thanks.

And last, but by no means least, I must thank my wife Robin. Her patience, her support, and her love have been the essential foundation for whatever I have accomplished during my years as a student at the University of Ottawa. Without her, I can say quite truthfully, I would not be where I am today.

Table of Contents

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

CHAPTER 1

INTRODUCTION

In a world where testing has become a pervasive fact of life, the question of test bias, particularly bias that adversely affects minority groups, has acquired considerable social and political significance. Test bias can cause a given group undergoing a test-based selection procedure to have its future success underpredicted. Since such a selection procedure would therefore tend to result in a lower selection rate for that group, the use of such a biased test in a selection procedure can be described as "unfair".

However, even tests applied correctly, i.e. in a non-discriminatory fashion, can still be affected by a different kind of bias against some groups. This second kind of bias has been termed item bias, because it pertains to certain individual items on a given test, rather than to the test as a whole. Though item bias can cause a test to be biased, it is not necessarily pervasive; it does not necessarily affect the entire test. A test that contains biased items can therefore be salvaged if the biased items are detected and then either corrected or removed.

Two methods have been employed in recent years to detect item bias, judgmental and statistical. Judgmental methods consist of the employment of one or more "expert judges" to assess items in a test in order to determine if there is a likelihood that a given group will find those items more difficult than another group because of something other than the trait being measured. The problem with judgmental methods is obvious; the judges may be "expert", assuming that such a rating can even be agreed upon, but they are still human and therefore

subjective (Ibrahim, 1992). While judgmental methods are widely employed in order to screen items for bias, it is still considered necessary to complement the judges' ratings with more objective statistical methods. Unfortunately, judgmental and statistical methods often give different results (Hambleton & Jones, 1992; Sudweeks & Tolman, 1990).

Even though statistical methods may indeed be more "objective" than judgmental methods, they cannot be said to actually detect item bias at all. Simply because a statistical procedure flags a given item as being "biased" does not mean that bias is necessarily present. A judgment of bias presupposes an understanding of the specific effect an item has on group performance. A judgment must still be made as to how the statistically discovered "bias" actually does affect a given group's performance on a test. For example, an item may be flagged as "biased", and yet no expert judges can determine just how bias is present. For this reason, when statistical bias detection methods are employed, the term "biased" is not used to describe items flagged by these procedures. Instead, the term "differential item performance" or "differential item functioning" is used. These terms do not contain the implication that bias is necessarily present, since they are simply descriptions of a statistical fact; the item in question functions differently for given groups taking a test.

Differential item functioning (referred to henceforth as dif) is what happens when one group of examinees finds an item more difficult than does another group of examinees of like ability. The focal group (FG) is the group of concern, with the reference group (RG) being the standard against which FG performance is compared. Classical test theory has been the source of some methods for the detection of dif. For example, the difficulty level of a given item for two or more groups can be compared. If a significant difference in difficulty level is found,

then dif may possibly have occurred. However, such a comparison fails to take into account the possibility that the test-takers in the groups being compared may have differing levels of ability. If one group is more able than another, then a difference in the difficulty level between the groups may not be evidence of bias at all, but is actually the result of the item doing its job--reflecting different levels of ability in performance on the item. This kind of difference has been termed "item impact" (Ackerman, 1992) to distinguish it from item bias. Approaches that do not account for the possiblity of item impact are termed unconditional because performance on the item is not conditioned on ability.

This disadvantage can be eliminated by the employment of conditional methods. In conditional methods, the examinees are matched on ability levels before their performance on an item is compared. Therefore, comparisons of difficulty are made only between examinees of the same ability level, even though they belong to different groups. Any performance differences between groups can therefore be attributed to dif with a higher degree of confidence than with unconditional methods.

One of the most often used conditional methods is the Mantel-Haenszel procedure. The Mantel-Haenszel (MH) is one of the chi-square based family of indices. It was originally developed by Mantel and Haenszel (1959) for the purpose of studying disease, but was modified by Holland and Thayer (1986) for application to dif research.

Holland and Thayer argued that the MH has several advantages over other forms of dif indices such as 3-parameter IRT procedures. The MH is cost effective, uses less involved statistical procedures than the "theoretically preferred" (Shephard, Camilli, & Williams, 1985) 3-parameter IRT model, and has a kind of intuitive obviousness that researchers can find

appealing. IRT based procedures, on the other hand, have several disadvantages. Most have no associated test of significance, need large sample sizes to work properly, require that stringent assumptions be made by the researcher, and are expensive and time consuming to run on computer. The MH, on the other hand, is cheap and easy to run on computer, and has an associated test of significance, the Mantel-Haenszel Chi Square (MHCHISQ). The MH procedure has become widely employed, and is used by such organizations as the Educational Testing Service to screen items for dif.

Recently, Swaminathan and Rogers (1990) have developed a procedure for the detection of dif based on the logistic regression (LR) model. The LR technique, like the MH procedure, is cheaper and easier to implement than IRT-based methods, according to Swaminathan and Rogers. Like the MH, the LR has an associated test of significance; unlike the MH, it can detect both uniform and non-uniform dif. Uniform dif occurs when the probability of answering an item correctly is uniformly greater for one group than the other across all ability levels. Non-uniform dif occurs when the difference in the probability of answering a given item correctly is not the same for each group across all ability levels.

What follows is a description of the MH and LR procedures.

The Mantel-Haenszel Procedure

When employing the MH procedure to test for dif, the first step is to establish a criterion for ability. Total test score is the most often employed criterion, with test scores being organized into score intervals or ability groups for both RG and FG members. Each item is

analyzed for dif, with the item under consideration being termed the studied item. For each ability level, 2x2 contingency tables can be constructed as shown in 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
RG	A_j	B_j	n_{Rj}
FG	C_j	D_j	n_{Fj}
Total	m_{1j}	m_{0j}	T_j

where 1 denotes a correct answer, 0 an incorrect answer; A_j and C_j denote the number of members of RG and FG respectively in the jth score interval who answered correctly; B_j and D_j denote the number of members of RG and FG in the jth score interval who answered incorrectly; n_{Rj} and n_{Fj} denote the total number of members of RG and FG, respectively, in the jth score interval; m_{1j} and m_{0j} denote the number of examinees in the jth score interval who answered correctly and incorrectly, respectively; and T_j is the total number of examinees in the jth score interval.

The ratio of the odds of success for RG and FG can be calculated from the tables for each score interval, weighted according to the number of examinees in each interval. This

produces the α_{MH}, defined by Holland and Thayer (1986) as the common odds ratio of success of the two groups across all ability levels. The α_{MH} can be defined as:

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

The α_{MH} can range in value from 0 to infinity, with $\alpha_{MH} = 1$ representing the null value, i.e., no dif. The α_{MH} is the average amount by which the odds that a RG member is correct on the studied item exceed the odds that a FG member with the same ability will also be correct on the item. When $\alpha_{MH} > 1$, the RG performs better on average on the studied item, and when $\alpha_{MH} < 1$, the FG performs better on the studied item than the RG.

The hypotheses associated with α_{MH} can be expressed as follows:

$$H_0 : \frac{p_{Rj}/q_{Rj}}{p_{Fj}/q_{Fj}} = 1 \text{ for all } j\text{'s} \qquad (2)$$

versus

$$H_1 : \frac{p_{Rj}/q_{Rj}}{p_{Fj}/q_{Fj}} \neq 1 \text{ for all } j\text{'s} \qquad (3)$$

where p_{Rj} and q_{Rj} represent the probability of a member of RG at the jth ability level answering the item correctly or incorrectly, respectively; and p_{Fj} and q_{Fj} denote the corresponding probabilities for the FG.

The hypotheses have an associated chi-square test of significance, the MHCHISQ, which can be written as follows:

$$MHCHISQ = \frac{(|\Sigma A_j - \Sigma E(A_j)| - 1/2)^2}{\Sigma Var(A_j)} \tag{4}$$

where

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

and

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

A significant MHCHISQ result indicates that the studied item is exhibiting dif. Under the null hypothesis, the MHCHISQ has an approximate chi-square distribution with one degree of freedom. Since the alternative hypothesis is non-directional, the MHCHISQ can identify dif that favours either the RG or FG. Holland and Thayer (1986) have claimed that the MHCHISQ is the uniformly most powerful unbiased test of the null hypothesis versus the alternative.

It is useful, given the presence of dif, to know the magnitude and direction of the effect. This information is provided by the Δ_{MH}. The Δ_{MH} is derived by performing a log transformation of α_{MH} in order to place it on the ETS difficulty scale. The ETS scale is created by a transformation of the item difficulty (as measured by the proportion of correct responses).

The item difficulty (p) measure is normalized by converting to z scores that correspond to the (1-p)th percentile. The z scores are then adjusted by means of a linear transformation to delta values which have a mean of 13 and and a standard deviation of 4. The Δ_{MH} is the logarithmic transformation of α_{MH} as follows:

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

The Δ_{MH} measures the average amount more difficult a member of the RG found the studied item than did a member of the FG of comparable ability. A negative Δ_{MH} indicates that the item in question was more difficult for the FG, while a positive Δ_{MH} indicates that the item was easier for the FG. A Δ_{MH} of 1 indicates a difference in item difficulty 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.

The z-score distribution has a mean of zero and standard deviation of one. A comparison of distributions causes the means to cancel, while the standard deviations can be expressed as a ratio. Since the delta difficulty scale has a standard deviation of four, the delta difficulty scale and a z-score scale have standard deviations in the ratio of four to one. The MH delta can therefore be expressed in terms of a scale of z-scores by simply dividing by four, which produces the MH-Z:

$$MH\text{-}Z = - 1 / 1.7 \ln (\alpha_{MH}) \quad (8)$$

Like the Δ_{MH}, the MH-Z has a value of 0 under the null hypothesis and gives a measure of both the direction and magnitude of dif in the studied item. The Δ_{MH} provides a measure in terms of the ETS delta difficulty scale, while the MH-Z does so in terms of a z-scale. Thus an MH-Z of .25 indicates a difference in item difficulty of about .10

The Logistic Regression (LR) Procedure

The standard logistic regression model is employed to predict from independent variables the probability of a correct response to an item. It is expressed by the following general equation:

$$P(u = 1|\theta) = \frac{e^{(\beta_0 + \beta_1 \theta_.)}}{(1 + e^{(\beta_0 + \beta_1 \theta)})} \qquad (9)$$

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

Swaminathan and Rogers (1990) modified the general LR equation in order to model dif as follows:

$$P(u_{ij} = 1|\theta_{ij} = \frac{e^{(\beta_{0j} + \beta_{1j} \theta_{ij})}}{(1 + e^{(\beta_{0j} + \beta_{1j} \theta_{ij})})}, \; i=1,\ldots,n_j \qquad (10)$$

where u_{ij} is the response of an individual i in group j to the studied item; θij is the observed ability of individual i in group j; β_{0j} is the intercept parameter for group j; and β_{1j} is the slope parameter for group j.

The model is used in order to predict the probability of a correct response when examinee ability and group membership are known. An absence of dif is indicated by equal logisitic regression curves, which occur when $\beta_{01}=\beta_{02}$, i.e., the intercepts are equal, and when $\beta_{11}=\beta_{12}$, i.e., the slopes are equal. When the intercept parameters are unequal ($\beta_{01} \neq \beta_{02}$), but the slope parameters are equal ($\beta_{11} = \beta_{12}$), the curves are parallel but do not intersect, and uniform dif is present. When the slope parameters are unequal ($\beta_{11} \neq \beta_{12}$), whether or not the intercepts are themselves equal, non-uniform dif is present.

The LR procedure as proposed by Swaminathan and Rogers (1990) uses a chi-square test of significance to test the null hypothesis of no dif:

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

The resulting statistic is a chi-square with two degrees of freedom. When the above test is used, the LR does not differentiate between uniform and non-uniform dif. If the LR procedure is used to test separately both uniform and non-uniform dif simultaneously, two separate chi-squares result, each with one degree of freedom.

While the LR procedure is iterative in nature and, therefore, moderately more expensive than the MH in terms of computer time, it has the advantage over the MH of being able to

detect both uniform and non-uniform dif. Like the MH, but unlike IRT based procedures, the LR has an associated test of significance, an important benefit given the difficulties of establishing cutoff points in the IRT model. However, the MH provides a measure of both the direction and magnitude of dif, which the LR model does not.

There have been many studies in which the performance of the MH procedure has been examined, while somewhat fewer studies have been focused on the LR procedure. In recent years, several researchers have compared the MH and LR procedures, both in terms of their distributional properties and their dif detection abilities. A review of this research follows.

CHAPTER 2

LITERATURE REVIEW

Since the MH procedure was first proposed by Holland and Thayer (1986), it has been compared to IRT-based methods (Baghi & Ferrara, 1989, 1990; Hambleton & Rogers, 1988; Hambleton, Rogers, & Arrasmith, 1988; Linacre & Wright, 1987; Sykes & Fitzpatrick, 1990) as well as to a variety of other dif detection procedures (Schulz & Geisinger, 1992). The LR procedure has not been as extensively examined, but has been compared to the MH in a number of studies (Brown, 1992; Ibrahim, 1992; Ocheing, 1992; Pang & Boss, 1993; Pang, Tian & Boss, 1994; Tian, Pang & Boss, 1994a, 1994b; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). In the following review some of the studies that have compared the MH to other dif detection procedures are examined. This is done in order to determine if the MH is an acceptable alternative to other established procedures. Then, studies in which the performance of the MH was examined are discussed in order to determine the strengths and weaknesses of the MH in relation to a variety of factors, including sample size, levels of item discrimination and item difficulty, etc. Then, studies where the performance of the LR procedures was examined are discussed, with most of these studies being comparisons of the relative performance of the LR and MH procedures under a variety of conditions.

The MH statistics have been compared to a variety of other well-established procedures for detecting dif. Hambleton and Rogers (1988) compared the IRT Area method to the MH procedure, using four samples of 1000 examinees each; two of the samples were Anglo-

American, and two were Native American. Seventy five of the original 150 items on the New Mexico High School Proficiency Exam (NMHSPE) were studied, with very easy items (p greater than .90) and poorly discriminating items (biserial correlation less than .10) being dropped because they can lead to difficulty in IRT parameter estimation. IRT 3-parameter models were then fitted to the samples separately.

Two analyses were performed, the first on all four of the samples, with a second analysis being conducted on samples of 650 each of Anglo and Native Americans in order to enable the researchers to examine the consistency with which each statistic flagged dif items across score distributions. In the second analysis, the ability range over which dif was calculated was restricted to two standard deviations above and below the mean test score for the Native Americans, because most Native American examinees were located in this range on the ability scale.

The first analysis resulted in the IRT Area method being consistent across samples about 73% of the time, while the MH was consistent about 80% of the time. Fourteen items were identified twice by the IRT method, nine by the MH, and seven of these items were common. Of the items identified by the IRT Area method, but not identified by the MH, most exhibited non-uniform dif. In the second analysis, Hambleton and Rogers found that matching the groups according to test score did not have a judgmentally significant effect on the results. The authors concluded that the IRT and Mantel-Haenszel methods agreed substantially in the items they identified as dif. In this and other studies, the MH procedure seems to have compared reasonably well to the "theoretically preferred" (Shepard, Camilli & Williams, 1985) IRT model.

Hambleton, Rogers, and Arrasmith (1988) compared the MH with three IRT-based

methods, the b-value plot method, the root mean square method, and the total area method. The data consisted of the responses of 937 Cleveland ninth graders to the first 76 of the 92 items on the Cleveland Reading Competency Test. Item parameters and ability estimates were obtained from the sample using the 3-parameter logistic model; then simulated item responses were generated using the same parameters so as to be consistent with the sample of responses of the students. These simulated data were used to establish cutoff points in order to permit the interpretation of IRT statistics so as to correspond to an alpha level of .01 for the MH.

The authors found that the MH detected six items at a significance level of .01. The number of items flagged ranged from five to eight across the four methods. The methods had a moderate level of agreement for which items exhibited dif, but the authors found that when certain methodological problems (difficulties in establishing cutoff points, Type I error, and poor item parameter estimates) were taken into account, the level of agreement among the four methods increased. The authors concluded that the MH and IRT methods agreed in the identification of dif items, but that the MH was preferable because it is quicker and cheaper.

Hambleton and Jones (1992) compared judgmental and statistical methods on their agreement in detection of dif items. They analyzed data from a statewide proficiency test, using a sample of 8,000 Anglo-American and 2000 Native American examinees. An item bias review form constructed by the authors was given to a group of expert judges to review 75 items from the test; all items of low discrimination (point biserial correlation less than .10) and unusual difficulty (p greater than .90 or less than .10) were not included. IRT area values were calculated across the range from the lower group mean minus three standard deviations to the upper group mean plus three standard deviations. The MH was applied to a random sample of

2000 Anglo-American and 2000 Native examinees, with each sample being subdivided further into two groups of 1000 each.

The cutoff for the MH was established at a significance level of .01, and the IRT area cutoff was set at .468. The authors found that the MH and IRT procedures were not in agreement in identifying items with dif, but did agree substantially on which items exhibited dif when they were used to identify only those items that had already been flagged as having dif by expert judges.

The MH indices have been compared to many other procedures for identifying dif as well (Camilli & Smith, 1990; Ackerman & Evans, 1992; Kromrey & Parshall, 1991). In general, it appears that the MH procedure compared either adequately or favorably to these other tests of dif.

The performance of the MH procedure itself has been examined by several researchers under a variety of conditions. Holland and Thayer (1986) argued that one of the major advantages of the MH procedure over IRT based approaches was that the MH was effective even with small sample sizes. Mazor, Clauser and Hambleton (1991) studied the effect of sample size on the MH statistics. They generated data using the 3 parameter IRT model. The data formed 3 sets of 2000 examinees each, with each set being normally distributed. Ability measures for the two groups were set to have a mean of 0 and a standard deviation of 1, and were designated Reference Group 1 (RG1), and Focal Group 1 (FG1). The third group had a mean of -1 and standard deviation of 1, and was designated Focal Group 2 (FG2). RG1 and FG1 were used to compare groups of equal ability, and RG1 and FG2 to compare groups of unequal ability.

Five different tests were generated, with each test having a common group of 59 non-dif

items. 80 dif items were also generated, and were split into five groups of 16 items each. Each of these groups was joined with the common pool of 59 items, forming five tests of 75 items each. Eighty items were formed from combinations of four levels of "a" (.25, .60, .90, 1.25), five levels of "b" for RG1 (-2.5, -1.0, 0, 1.0, 2.5), and four levels of difference in "b" between RG and FGs (.25, .50, 1.00 and 1.50). The "c" parameter was fixed at .20. The levels of "a" and differences in "b" were completely crossed for each of five 16 item groups, each with one of the five "b" balues.

The MH procedure was run for each test comparison, first using all 2000 examinees in each group, then 1000, 500, 200, and 100 examinees. To minimize the effect of chance variability in the smaller groups, the run with 500 examinees was replicated once for each set, while the samples of 200 and 100 were replicated twice.

The results showed that with sample sizes of 100 to 200, the number of dif items correctly identified was "very small". With a sample size of 500, more than half of the dif items were missed. With a sample size of 1000, the MH correctly identified as dif 61% of the dif items when the ability distributions were equal, and 58% when the distributions were unequal. At a sample size of 2000, 74% of dif items were correctly identified when ability distributions were equal, and 64% with unequal distributions.

The items most likely to be missed at even large sample sizes were those of high difficulty, low discrimination, and small differences in "b" between groups. All of these trends increased as sample size decreased. However, Mazor et al. found that the MH was adversely affected by low item discrimination ("a" = .25) even at large sample sizes. It was also found that unequal ability distributions between FG and RG made the MH less sensitive to dif across all

sample sizes.

Differences in item difficulty were analyzed at sample size 2000. When ability distributions were equal, the largest p-value difference missed was .04, and the smallest difference identified was .02. At sample size of 1000 and unequal distributions, the largest p difference missed was .08 and the smallest identified was .03. At sample sizes of 500, 200, and 100, the largest differences in p-value missed were .08, .17, and .23 respectively, and the smallest p differences identified were .07, .07, and .15 respectively. However, when the ability distributions were not equal, the largest p differences missed were higher: .07, .15, .17, .23, and .29 at sample sizes of 2000, 1000, 500, 200, and 100 respectively.

Mazor et al. did not find the relatively poor performance of the MH at small sample sizes surprising, since any statistic becomes less powerful the smaller the sample size. However, a large proportion of dif items were missed even at a sample size of 2000, leading the authors to conclude that the MH statistic should be employed with caution, especially at sample sizes less than 1000.

Clauser, Mazor, and Hambleton (1991a) examined the effect of item discrimination and item difficulty on the MH. The study was focused on the statistical characteristics of items with dif which go undetected by the MH procedure. One of the questions the study was designed to answer was the effect of "a" values on dif detection when the "a" is equal for RG and FG, but high for some items and low for others.

Data were simulated as in the Mazor et al. study cited above, with five 75-item tests being simulated, and with "a", "b" and "c" parameters as above. Examinee responses were held constant for the 59 common items. There were 1000 examinees in each group, and ability

distributions were generated so as to be equal between the groups (mean=0, SD=1). The simulations were then repeated but with mean ability for FG of -1.00 and for RG of 0.00.

For groups of equal ability, the MH identified 49 of the 80 dif items, and misidentified 1 of the 59 non-dif items. An increase in the level of item discrimination produced a significant effect on the MH, causing the probability of an item being detected for dif to rise substantially. Low levels of "a" produced low MHCHISQ values, even when the difference in "b" was large for the item. This led Clauser et al. to conclude that the MH would tend to lead to the elimination of high "a" items from a test, while dif items with low "a" might still go undetected. When "b" itself was very high (around +2.50), the MH did not detect dif items.

Differences between RG and FG on "b" had an effect on the probability of dif items being flagged, with the probability rising substantially as "b" differences increased (hardly surprising, since difference in "b" was how dif was defined). As mentioned above, "b" differences had an effect on the probability of dif detection when item discrimination was high.

It was concluded that the MH does not function equally well across the range of the "b" scale. The MH was "essentially blind" to dif in high difficulty items. This fact is not surprising, since high "b" items can perform differentially for only the few members of high ability cells. Since the MH is weighted by the number of examinees in each ability cell, it tends to miss dif when cell group populations are sparse. Clauser et al. recommended that if a test is being designed to identify a small number of high ability examinees, then the MH should be run on a selected group of scores from the most competent examinees if it is to function optimally in the detection of dif.

Clauser, Mazor, and Hambleton (1991b) investigated the influence of the criterion

variable on the MH. They argued that when total test score is used as the criterion variable for the MH, the user is assuming that the test is reliable and valid. However, even tests that appear to be unidimensional may include items that measure more than one skill. If a test is multidimensional, then different groupings of test items analyzed might produce different MH results, rendering doubtful any conclusions about dif derived from the application of the MH (this is true of any dif detection procedure, of course).

To examine this potential problem, Clauser et al. analyzed data from the 1982 administration of the New Mexico High School Proficiency Exam (the same data set was used by Hambleton and Rogers (1988), and Hambleton and Jones (1992)). From the overall test taking population, two random samples of 1000 Anglo-American and 1000 Native American were drawn. Low discriminating (biserial correlation less than .10) and very easy (p greater than .90) items were removed from the original pool of 150 items, leaving 91 items in total. These items were divided into three groups of 75 items each, with each item being present in at least 2 of the 3 groups. The MH was calculated and dif items were identified, with items considered to exhibit dif only if they were identified as such in every group of which they were a member. The items were then categorized into four subtests (these tests were not the same as the original test subtests). These subtests (Math, Reading, Prior Knowledge, and Charts) were analyzed again with the MH. For purposes of control, the 91 items were then randomly assigned to three groups of 30, 31 and 30 items, and the MH was rerun. This run was performed in order to determine the degree to which the MH was affected by the reduction in the number of items being analyzed in the four subtests.

The MH identified 22 of the 91 items as dif on its initial run. When the items were

grouped into the four subtests, the MH ceased to identify 7 of the 22 items as dif, while 11 new items were identified as dif. The random groups of 30, 31 and 30 items were then analyzed, and only minor changes in dif detection were found. Several of the 91 items analyzed were of the "No Clear Best Answer" type (NCBA). The authors argued that these items required special knowledge to answer and would otherwise be seen as ambiguous. This meant that these items could be seen as truly culturally biased. Six of the twelve items of this type were flagged as dif in the 75-item runs and remained so in subsequent subtest runs (100% agreement). This was not the case with items with a clearly best answer.

The authors concluded that since 32% of the items flagged on the first run ceased exhibiting dif in the subtest runs, the grouping of test items had an effect on the MH and should therefore be considered by test makers as a variable. They also found a significant effect for test length on the MH, which identified more items as dif when test length decreased, even for the randomized control subtests. Because test length had an effect on the MH, the issue of dimensionality was not resolved by the study, since any subtest effects could have been confounded with test length.

Ryan (1990) examined the stability of the MH across samples of test takers and at different sample sizes. She collected data from the administration of the Second International Mathematics Study. The overall number of subjects was 670 Black and 5015 White examinees. The item pool from which the test was constructed consisted of 64 arithmetic items, 42 algebra items, 42 geometry items, 26 measurement items, and 18 statistics items. Eight items from each of the five content areas were grouped together to make a 40-item core test which was taken by all examinees. The rest of the items were randomly assigned to create four different 35-item

rotated tests, each of which was combined with the 40 item core to create four different 75-item tests. These tests were assigned randomly to one quarter of the sample, so each examinee completed a common pool of 40 items and one of four different groups of 35 items.

Ryan conducted a baseline analysis using 670 Whites to form a FG, with the rest of the Whites making up the RG. Ryan called this WW (White-White). The sample was analyzed using the MH delta and the MHCHISQ on the 40 core items. Four smaller samples were formed by dividing the 670 White examinees in the Set 1 FG to form four FGs ranging in size from 151 to 166, and by dividing the Set 1 RG into quarters to create four RGs ranging in size from 1025 to 1064. These FGs and RGs were paired to create four samples called WW1-WW4. These four samples were also analyzed on the 40-item core test using the MH delta and the MHCHISQ. The four samples together with the larger sample from which they came made up Set 1 of Ryan's study. Set 2 was formed with 670 Black examinees making up the FG and all 5015 White examinees making up the RG. As in Set 1, the large sample (called BW) was analyzed on the 40-item core test using the MH delta and MHCHISQ, and four smaller samples (BW1-BW4) were also analyzed. In both sets, the total test score functioned as the criterion variable. Various descriptive statistics were analyzed, including mean values and standard deviations for the MHCHISQ and MH delta.

Ryan found that the mean MHCHISQ value for the WW sample was 1.1, while that of the BW sample was 6.6, while the standard deviations were 1.6 and 8.9 respectively. The mean MHCHISQ values for the smaller samples ranged from .75 to 1.2 for WW1-4, and from 1.8 to 3.3 for BW1-4. The standard deviations for WW1-4 ranged from .9 to 2.1, and for BW1-4 from 2.4 to 4.6. MHCHISQ was significant for the BW samples about half the time, and not

at all for the WW samples, which was expected since the White students were from the same examinee group, even when they were divided into an RG and FG for purposes of a baseline analysis. The MH delta had a mean value for the WW sample of .005 and a standard deviation of .26, while the BW sample had a mean MH delta of -.017 and a standard deviation of .65. The means for WW1-WW4 ranged from .002 to .012, with standard deviations ranging from .48 to .56. The means for BW1-BW4 ranged from a low of -.003 to .036 and standard deviations ranged from .75 to .91.

Ryan also analyzed the MHCHISQ and MH delta by correlating their values for the small WW samples with one another, the small BW samples with one another, and each of the smaller samples with the larger sample from which they came. She found that the correlations of the MH delta for the small WW samples were low (-.13 to .26), while the correlations of the MH deltas for each of the small samples with the larger WW sample were larger. Ryan cautioned, however, that the latter correlations were probably spurious, given the overlapping of the small samples with the larger sample. She found that with the BW samples the correlations of the MH delta were larger than for the WW sample (.74 to .88) and the same was true for the MHCHISQ (.42 to .66).

Ryan concluded that the low correlations for MH statistics when sample sizes were small indicated that sample size is directly related to the effectiveness of the MH. She also argued that the large differences between Black and White examinees in mean scores on the test may have reduced the stability of the MHCHISQ, indicating that the MHCHISQ is sensitive to unequal ability distributions.

Gutierrez (1988) studied the performance of the MH-Z under controlled conditions of the

null hypothesis. She examined the effect of sample size, the "a" parameter, and the "b" parameter on the MH, including the effect of the "b" value being extremely high or low. She also examined the effect of these variables on the cutoff values for the MH-Z, as well as the relationship of false positive identifications to those cutoffs. A 40 item test was generated, with "b" values grouped into three ranges (-2.0 to -.8, -.7 to .6, and .7 to 2.0), "a" values of .7, 1.0, and 1.3, and sample sizes of 450/150, 900/300, and 1350/450 (RG to FG). Ability scores were randomly generated from a normal (0, 1) distribution, and cutoffs were computed for $P_{2.5}$, P_5, P_{95},and $P_{97.5}$. One hundred replications were performed, and 3x3 ANOVAs, sample size by "a", were conducted on the standard deviation of the MH-Z across the forty items of the test. As well, a repeated measures ANOVA was carried out on the standard deviation of the MH-Z across the 3 levels of "b", and a MANOVA was conducted on the cutoff values for the four computed percentiles.

Gutierrez found that the sample size by "a" interaction was significant at the .05 level. As expected under the null hypothesis, increasing the sample size decreased the standard deviation of the index, and increasing the "a" inflated the standard deviation. The repeated measures ANOVA led to the finding that the sample size by "a" by "b" interaction had a significant effect on the standard deviation. As "a" increased, the standard deviation also increased, but when the "b" values were close to the centre of the distribution, the standard deviation got smaller, and this interaction was even more pronounced with small sample sizes and large "a".

Gutierrez also found that the cutoffs decreased with an increase in sample size, increased with "a", and were additively inflated when sample sizes were small and "a" large. The number

of false positive identifications decreased when sample size grew, while an increase in the value of "a" produced an increase in false postives. When values of "b" were extreme, more false positives occurred than when when "b" values were near the centre of the distribution. Gutierrez concluded that since all of the variables she studied had a significant effect on the distribution of the MH-Z, the MH-Z was clearly influenced by variables other than dif.

Since the LR procedure was first described by Swaminathan and Rogers (1990) several researchers have compared the MH and LR procedures. Swaminathan and Rogers (1990) compared the MH and LR procedures on their consistency across sample size, test length, and type of dif (uniform/non-uniform). They generated data with sample sizes of 250 and 500, differing test lengths (40 items, 60 items, 80 items), and with differing types of dif (uniform and non-uniform). The two procedures were compared on the percentage of uniform and non-uniform dif items correctly identified, and on the percentage of non-dif items incorrectly identified (Type I error). Twenty replications were run on the combination of the sample size of 500 and test length of 80 items, while no replications were run under any of the other conditions.

Swaminathan and Rogers found that the MH and the LR procedures were about equally likely to detect uniform dif, but were both more successful at detecting dif at the larger sample size of 500 (75% success at 250, 100% success at 500). The MH procedure was completely unable to detect non-uniform dif at either sample size, while the LR procedure detected non-uniform dif about half the time at sample size 250 with short test length, and about 75% of the time at sample size 500 with longer test length. The MH and LR procedures seemed equally sensitive to sample size, providing they were being employed to detect uniform dif.

Ochieng (1992) examined the effects of sample size, item difficulty, item discrimination, and ability distribution on the MH statistics and the LR procedures under the null hypothesis. He analyzed the distribution of MHCHISQ, Δ_{MH}, LRU (the logistic regression index for uniform dif), and LRN (the logistic regression index for non-uniform dif), using as dependent variables the mean, standard deviation, skewness, kurtosis, and P_{90} and P_{95}.

Simulated data were generated under the null hypothesis, meaning that no actual dif was present in the items generated. One hundred replications were made for each condition for each dif index.

The simulated test had 63 items. Item response strings were generated under a 3-parameter IRT model with the "c" parameter fixed at 0.15, the "a" parameter set at three levels (0.6, 1.0, and 1.4), and the "b" parameter set at 21 levels ranging from -2.0 to +2.0, with intervals of .20. Three different sample sizes were simulated: 400, 800, and 1600, with the ratio of FG to RG within each sample set at 1:3. Ability distribution had two levels: one level where the FG and RG were of equal ability (mean equal to 0, standard deviation of 1); and a second level where the FG and RG were of unequal ability (RG mean of 0, SD of 1; FG mean of -0.5, SD of 0.83).

Ochieng found that the LRU was not affected by sample size. However, he did find that sample size had significant effects on the MHCHISQ, the Δ_{MH}, and the LRN. In addition, Ochieng found that the LRU was not affected by varying levels of "a", while the mean of the MHCHISQ was significantly affected, as were the mean and standard deviation of the Δ_{MH}. In each case, higher levels of "a" were directly related to the values of the mean and/or standard deviation. Ochieng also found that the cutoffs P_{90} and P_{95} for the ΔMH were significantly

affected by level of "a", as were the cutoffs for the LRN. When high discrimination was present and sample sizes were small, the P_{90} of the MHCHISQ was smaller than expected, suggesting that the use of tabled values would lead to a decrease in detection of dif items.

Ochieng found that the "b" value had no significant effect on the LRU, while levels of "b" did have a significant effect on the mean and standard deviation of the LRN. Levels of "b" also significantly effected the P_{90} of the MHCHISQ. Ochieng found that ability distribution had no significant effect on the MHCHISQ, Δ_{MH}, LRU, and LRN.

Ibrahim (1992) examined a variety of dif detection procedures, including the MHCHISQ, the Δ_{MH}, the LRU, and the LRN. Item response strings were simulated under the 3-parameter IRT model, with "c" fixed at .15. Examinee samples were generated from normal ability distributions. There were two test lengths (42 and 66 items), and 4 arrangements of "a" values. In arrangement D1, all the "a" values were .80, in arrangement D2, the first half of the test had "a"=.80, the second half "a"=1.20. In D3, the first half had "a"=1.20, the second half had "a"=.80. In D4, all items had "a"=1.20. "B" values ranged from -2.0 to 2.0, and were grouped into "low", "medium", and "high" "b" items. The dif items, 6 in all, were generated by manipulating either "b" or "a" values for the FG, with 3 of the items showing uniform dif, 3 non-uniform dif. Sample sizes were in the ratio of 8:3 (RG to FG), and had 3 levels: 1600/600, 800/300, and 400/150. One hundred replications were run under each experimental condition, with 3 percentiles (P_{90}, P_{95}, and P_{99}) being computed for each dif index within each condition. MANOVAs and posthoc ANOVAs were conducted to test for differences in mean percentile values in each test across conditions.

Ibrahim found that under the null hypothesis sample size had a significant effect on the

means of the percentiles of the MHCHISQ, Δ_{MH}, and the LRN, but not on the LRU. The dif detection rate of the MHCHISQ was directly related to sample size. For non-uniform dif, the detection rates were .68, .57. and .42 for the samples sizes as they ranged from large to small respectively. For uniform dif, the MHCHISQ had high (greater than .91) detection rates at all 3 sample sizes. The MHCHISQ also tended to produce more false positives at larger sample sizes.

The LRU and LRN detected larger proportions of dif items at large sample sizes, with false positive rates being approximately equal at all sample sizes (slightly larger at 1600/600). Ibrahim also found that item discrimination had no significant effect on the dif detection rates of the MHCHISQ, the LRU, or the LRN.

Ibrahim found that the MHCHISQ's detection rates were negatively related to level of "b" when dif was uniform, and much lower when "b" levels were moderate and dif was non-uniform. When dif is non-uniform, the ICC curves tend to cross in the middle of the ability distribution, i.e., when "b" is moderate, which would result in the areas between the curves being both positive and negative, and therefore cancelling one another out. Since the MHCHISQ is a signed statistic (up to the point where the numerator is squared), it will tend to miss this kind of dif (Gutierrez, 1988). The MHCHISQ was able to pick up non-uniform dif when "b" values were either high or low, perhaps because dif may appear to be uniform at the extremes of the distribution. The false positive rates for the MHCHISQ over "b" values were .075, .111, and .094, respectively. The LR procedure had higher detection rates when "b" decreased, with the highest false positive rate occuring when "b" was moderate with the lowest false positive rate at low "b".

Brown (1992) examined real data from a Grade 4 reading assessment test in British Columbia. She compared the MH and the LR across sample size and over replications. The test was composed of two short subtests. Sixteen comparison groups were created by randomly selecting cases by gender and/or region across 6 sample sizes: 1000/1000, 750/750, 500/500, 300/300, 200/200, and 100/100. The MHCHISQ, LRU, and LRN were computed separately for each of the two subtests, and replications were run at each sample size one to five times. Detection rates were low (significance level of .05) for all three indices, with the LRU performing marginally better than the MH. The false positive rate was about 6% for all three procedures, with the detection rates being substantially higher for larger sample sizes.

Rogers and Swaminathan (1993) examined the distributional properties of the MH and the LR procedures and also compared them in terms of their relative power. In the distributional study, the authors constructed empirical sampling distributions under several conditions. Two independent variables were of prime importance in the study: sample size and goodness of fit between model and data. Two levels of model/data fit ("good" and "poor") were crossed with two levels of sample size (250 and 500). Data with "good" fit were generated using the two-parameter logistic IRT model, and data with "poor" fit were generated using the three-parameter logistic IRT model. Under the three parameter model, all items had "c" values of .2. Item parameters were chosen for a 40-item test based on real test data and were selected to produce an approximately normal distribution of scores. Item responses were generated so that items were unbiased. For each combination of model/data fit and sample size, 100 replications were performed. Five of the 40 items were chosen for study so as to vary in levels of "a" and "b". The parameters of the five items were as follows: low "b" and low "a" (b=-1.5, a=.6),

moderate "b" and moderate "a" (b=0, a=1), high "b" and high "a" (b=1.5, a=1.6), high "b" and low "a" (b=1.5, a=.6), and low "b" and high "a" (b=-1.5, a=1.6). The LR and MHCHISQ were calculated for each of these five items, and empirical sampling distributions were constructed.

In the power study, thirty two conditions were simulated by crossing two levels of model-data fit ("good" and "poor", simulated in the same way as in the distributional study), two levels of sample size (250 and 500), two levels of test length (40 items and 80 items), two levels of test score distribution shape (normal and negatively skewed), and two levels of percent of items with dif (15% including the item of interest and 0% other than in the item of interest). Uniform and non-uniform dif were simulated for each condition, and four sizes of dif were studied, corresponding to area values of .2, .4, .6 or .8. Uniform dif was simulated by keeping the "a" parameters for both groups the same but varying the "b" parameters, while non-uniform dif was simulated by varying the "a" parameter between groups and keeping the "b" parameters the same. A further group of "mixed non-uniform" items was generated by varying both "a" and "b" parameters between the groups. Thirty five dif items were constructed, and were added separately to the two tests so as to have 15% of the items exhibit dif whether the test had 40 or 80 items, or alternatively, so as to have only the item of interest having dif. Each of the thirty two conditions described above was replicated 20 times.

Rogers and Swaminathan found that the distributional assumptions of both procedures were met adequately, although the LR procedure fit the expected distribution less well with an item that was highly difficult and discriminating. They argued that this finding was of little practical significance on most ability tests, since such tests rarely contain items that are both

very difficult and highly discriminating. In the power study, both the MH and LR procedures were virtually equal in their ability to detect uniform dif, with the MH having a small advantage over the LR. However, the LR procedure detected 57% of the dif items when non-uniform dif was greater than .2 and the model-data fit was poor. The MH was unable to detect "strictly" (also called "disordinal") non-uniform dif, i.e., dif that resulted in ICCs crossing at the center of the ability distribution (Ibrahim, 1992). The authors found that the MH was not affected by the percent of items with dif and attributed this to the two-stage "purification" procedure associated with the MH. The authors concluded that the LR procedure provided a worthwhile alternative to the MH, given its superior sensitivity to non-uniform dif in tests of moderate difficulty.

Pang and Boss (1993) studied the effect of sample size, item discrimination, and item difficulty on the distribution of the LR indices under the null hypothesis. Three levels of sample size for RG and FG (250, 500, and 1000 each) were crossed with 3 levels of "a" (.7, 1.0, and 1.3) and 3 levels of "b" ("low": -2.0 to -.7; "medium": -.6 to .6; and "high": .7 to 2.0). LRU and LRN were calculated and 100 replications were performed under each condition. Mean, standard deviation, and skewness were computed for each of the LR distributions for each condition, and 3 cutoffs (P_{90}, P_{95}, and P_{99}) were computed for each condition. MANOVAs were conducted on each of the descriptive statistics.

The results showed that the LRU was not affected by sample size, although the mean for the LRU was slightly underestimated when compared with tabled values of a chi-square distribution. The LRN was not affected by sample size either, which contradicts the finding of Ibrahim (1992) and supports Ochieng (1992).

For LRU, sample size had no significant effect on the three cutoffs, although the critical values were slightly overestimated at P_{90} and P_{95} and underestimated at P_{99}. The overestimation at the two lower cutoff values might lead to a higher rate of false-positive identifications if tabled values are used as the criterion of identification.

Pang and Boss also found that the LRU was not affected by level of "a", but that the LRN cutoffs at P_{90} and P_{95} had higher values than expected when discrimination was high ("a"=1.30). They did find that the LRU was not affected by level of "b", but the mean of the LRN exceeded the expected value when "b" was high. As well, the standard deviation was underestimated when "b" was moderate or low.

Tian, Pang, and Boss (1994a) examined the consistency of the LR and MH procedures in dif detection across sample size and over replications. In addition, they investigated the effect of using total test score versus subtest score as the criterion variable. The data in the study were composed of the item responses of 183,356 Caucasian examinees from the 1989 administration of the English Test, Form 39B, of the American College Test (ACT). The English Test has 75 items and is divided into two subtests: a 40-item Usage/Mechanics test, and a 35-item Rhetorical Skills test. Of the examinees, 75,854 were male, and these were arbitrarily classified as the FG, while the 107,502 female examinees constituted the RG. Comparison groups were established with sample sizes of 100/100, 250/250, 500/500, 1000/1000, and 2000/2000. The MH and LR statistics were calculated with total test score as the criterion, and were then recalculated using subtest score as the criterion variable. Thirty replications were performed. Since real data cannot be said definitively to contain dif or not, a standard was developed by the researchers in order to classify items on degree and direction of dif. Based on the calculated

MH-Z values for all items at sample size of 2,000 and across 30 replications, items were classified as possible dif if the absolute mean MH-Z value was at least .15 but less than .20. Items were classed as probable dif if the absolute mean MH-Z value was between .20 and .25, and as definite dif if the absolute mean MH-Z value exceeded .25.

Tian et al. found that the standard deviation of the MH-Z tended to decrease as sample size decreased. Of the items found to have dif, five items were classed as possible dif, three as probable dif, and one as definite dif. Criterion variable did not appear to affect dif detection for either procedure. However, sample size did have an effect on the power of both the MH and LR statistics. When sample size was 100, the highest rate of dif identification was eight replications out of thirty. Even when mean effect size was large (MH-Z > .25), detection rates were low when sample size was small. However, when sample size was large (2000), detection rates were high even when mean effect size was small (MH-Z = .15). The researchers also found that the MH was able to detect non-uniform dif in a highly difficult item (p = .25), while it almost failed to indentify dif in a moderately difficult item (p = .49). Tian et al. argued that this effect occured because the FG and RG item characteristic curves for the highly difficult item crossed at the upper end of the ability spectrum, making the dif effectively uniform across most of the range of the ability scale and, therefore, detectable by the MH.

The researchers concluded that total test score versus subtest score had no effect on either the MH or LR procedures. However, sample size affected both the MH and the LR. The LR procedure detected dif somewhat more often, but also had a slightly higher false-positive rate. In general, Tian et al. concluded that the MH and LR statistics were interchangeable in their ability to detect uniform dif. They argued as well that a sample size of at least 500 to 1000 is

needed in order for the MH and LR procedures to perform reliably.

The effects of sample size and criterion variable on the MH and LR were investigated in a study by Pang, Tian, and Boss (1994). The data set was composed of the same examinees as in Tian et al. (1994a), but was based on scores on the Math test, which has three subtests: Elementary Algebra (Subtest 1) with 24 items, Intermediate Algebra/Coordinate Geometry (Subtest 2) with 18 items, and Plane Geometry (Subtest 3) with 18 items. Sample sizes were 100/100, 250/250, 500/500, and 1000/1000 for each procedure, and the criterion variable was total test score versus subtest score. Thirty replications were run, with males being the RG and and females the FG. MH and LR statistics were computed at both .05 and .01 levels of significance. Items were classified as possible, probable, and definite dif as in Tian et al. (1994a), using the absolute mean value of the MH-Z computed for sample size of 1000 as the classification criterion.

Pang et al. found the LR to be slightly more powerful at dif detection than the MH for uniform dif. They also found that using subtest score as the criterion variable generally lowered detection rates for both procedures, with some items becoming classified as less likely to be items with dif given the classification scheme used in the study.

The LR procedure was found to have a higher false positive rate under most of the conditions studied (the only exception being sample size of 100 at .01 level of significance). Although there were some exceptions, increasing sample size tended to increase false positive identifications by both procedures. As in Tian et al. (1994a), the fact that the LR had a higher false positive rate than the MH could account for the LR's generally higher detection rate overall.

Pang et al. concluded that the LR and MH are subject to similar sample size effects, and show the same tendency to have their performance adversely affected by the use of subtest scores as the criterion variable. They also concluded that a minimum sample size of 500 is necessary for either procedure to detect definite dif, with a sample size of 1000 needed in order to reliably detect probable and possible dif.

Tian, Pang, and Boss (1994b) studied the effects of sample size and criterion variable on the MH and LR using as their data set the item responses of 75,752 male and 107,256 female examinees from the 1989 administration of the Reading Test of the ACT (Form 39B). The Reading Test contains two subtests: a Reading Arts/Literature subtest of 20 items, and a Reading Social Studies/Science subtest, also with 20 items. The same methodology was used as in Pang et al., with sample sizes of 100/100, 250/250, 500/500, and 1000/1000, levels of significance at .01 and .05, total test score versus subtest score as the criterion variable, and thirty replications. As in Pang et al., absolute mean value of the MH-Z at sample size of 1000 was used to classify level of dif for each of the items.

As in Tian et al. (1994a), and Pang et al. (1994), sample size was observed to have a strong effect on the power of the MH and LR statistics, with detection rates being low at small sample sizes even with large mean effect sizes, and high at large sample sizes even when mean effect size is relatively small. Also consistent with the findings of Tian et al. (1994a), mean MH-Z values were generally consistent across sample sizes, but the standard deviation of the MH-Z increases as sample size decreased.

The effect of criterion variable on the MH and LR procedures may have been confounded by a speed factor, since the last five to ten items were not answered by a large number of

examinees. As well, the subtests employed may not have been unidimensional, causing unexpectedly large numbers of items to be flagged as dif. This finding contradicted Tian et al. (1994a), who found that dif detection rates were not affected by the use of subtest scores as the criterion variable, perhaps because the subtests employed in their study were unidimensional. The false positive rates of the MH and LR were very similar and were within the expected range.

Tian et al. concluded that while both procedures were influenced by variation in sample size, the MH appeared to be less susceptible to variation in sample size. The LR procedure was found to be slightly more powerful than the MH at detecting uniform dif. The hypotheses of uniform and non-uniform dif were tested in this study using two separate chi-square tests, each with one degree of freedom. The increase in power gained by the resulting conservation of one degree of freedom may have contributed to the LR's higher rate of dif detection. In general, the LR and MH procedures were equally effective in detecting dif, especially at sample sizes of 500 to 1000.

Summary and Research Questions

It appears that the MH and LR procedures are more successful at detecting dif at larger sample sizes (Mazor et al., 1991; Ibrahim, 1992; Swaminathan & Rogers, 1990), an unsurprising finding given that both the MH and LR have chi-square based tests of significance, and the power of chi-square tests is known to be inflated substantially by large sample sizes.

However, the answer to the question of what constitutes a small or large sample size has remained somewhat ambiguous. Dif detection by the MH was especially poor at very small sample sizes (Mazor et al., 1991; Ryan, 1990), while the power of both the MH and LR procedures was adequate at samples sizes of 1000 or more (Ibrahim, 1992; Brown, 1992). Some researchers have found that the detection rate of the LR was high at sample sizes of around 500 (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993), and it has been suggested (Tian et al., 1994a, 1994b) that a minimum sample size of between 500 and 1000 is necessary for the MH and LR procedures to produce reliable results.

The studies in which the effect of sample size on the distributions of the MH and LR was examined generally led to the conclusion that increasing sample size tended to affect the distributions of the MH and LR (Ibrahim, 1992; Ochieng, 1992), although some researchers (Pang & Boss, 1993) found the LRU to be unaffected as sample size increased.

However, in few studies has the effect of sample size on the MH and LR procedures been examined when effect size (i.e., the degree of dif) varies. Those studies that have examined this relationship (Pang et al., 1994; Tian et al., 1994a, 1994b) have either used examinee data from published tests, or have used simulated data with only a single replication (Clauser et al., 1991a; Mazor et al., 1991). The sole exception has been Rogers and Swaminathan (1993), and their study had only thirty replications.

Since one of the claimed advantages of the MH and LR procedures is their less stringent sample size requirements in comparison to IRT-based methods, it would be useful to examine thoroughly the performance of the MH and LR procedures at small sample sizes. Since IRT-based methods require sample sizes greater than 1000, and since the performance of the MH and

LR procedures appears to be adequate at sample sizes greater than or equal to 1000 (Ibrahim, 1992; Ocheing, 1992; Pang et al., 1994; Tian et al., 1994a, 1994b;), it is perhaps less important to examine the relative performance of the MH and LR at larger sample sizes.

However, the ability of either procedure to detect dif at any given sample size undoubtedly depends on the degree of dif present. A given sample size may be so small that even a large degree of dif would not be detected. Alternatively, while a low degree of dif might not be detectable at a certain sample size, a moderate or large degree of dif might be. Either result could be of considerable practical significance to test administrators, who may not be concerned if low dif items are present in a test, or who may wish to know the sample size at which dif is not detected at all, no matter how great the effect. To answer these questions adequately, effect size should be considered together with sample size.

The quality of the performance of a dif detection procedure is a function of its ability to detect dif when it is present, and to not detect dif when it is not present. Detection of non-existent dif is an instance of Type I error, or "false-positive" identification. Tian et al. (1994a, 1994b) and Pang et al. (1994) found that the LR procedure was slightly more powerful than the MH, but also had slightly higher false positive rates. However, these studies involved the use of real data. In real data, the presence and degree of dif is not exact using purely statistical means. This disadvantage is avoided when simulated data are used and the presence and degree of dif can be controlled item by item. By examining the detection and false positive rates of the two procedures on a set of simulated data, it may be possible to determine the degree to which the false positive rate of each procedure is related to its power. The MH and LR could then be compared on their dif detection rates and their false-positive rates across sample sizes and with

different levels of dif.

The proposed study is therefore focused on the following questions:

1. Which of the MH and LR procedures is more powerful in the identification of dif at sample sizes less than 1000?

2. What is the relationship of effect size to the power of the MH and LR at a given sample size? Is there a sample size at which dif is never detected, no matter what its degree, and if so, is it the same sample size for both procedures? What is the minumum sample size for reliable dif detection when the effect size is moderate? When it is large?

3. What is the relationship between the power of the MH and LR and their false-positive rates under the above conditions?

CHAPTER 3

METHODOLOGY

The methodology employed in the study is presented in this chapter. The data generation procedures are described, followed by descriptions of the characteristics of the samples and the item parameters used in the study. A power analysis follows, in which the measure of effect size employed in the study is described and related to the typology of Cohen (1992). The method of data analysis is then described.

Data Generation

In order to test fully the relationship between sample size and effect size, and in order to account for the influences of such factors as item difficulty and item discrimination, it is necessary to be able to know the precise values of each of these variables and to manipulate them with a high degree of control. This level of knowledge and control can only be achieved if the data are simulated, since knowledge of the degree and direction of dif in real data can never be exact. In this study, therefore, Monte Carlo methods were used to simulate data in which all the variables of interest, sample size, effect size, item difficulty, and item discrimination, were directly manipulated by the researcher.

The two studies cited in which the relationship of sample size and effect size to power was examined using simulated data (Clauser et al., 1991a; Mazor et al., 1991) only involved the

carrying out of a single replication. In order to provide a more reliable set of conclusions to be drawn concerning the research questions, 100 replications were carried out in this study.

Examinee samples were randomly generated with abilities from a normal (0,1) distribution, and item response strings were simulated so as to constitute the data set to be analyzed in the study. It was assumed that the use of dichotomous (0,1) scoring would not result in any loss of information, even though underlying traits are typically understood to be continuously distributed. The test used in the study included 66 items, which entailed the further assumption that 66 items were sufficient to measure the ability of interest. Since simulated data were used in the study, questions of the unidimensionality of the test and local independence between items were not a problem. Data were generated based on the three-parameter IRT model, using the DATAGEN program (Carlson, 1983).

Characteristics of the Samples

Sample size in the study were equal for RG and FG, and had five possible values: 100/100, 200/200, 400/400, 600/600, and 800/800. This range of values permitted a thorough examination of the performance of the MH and LR at sample sizes less than 1000.

Characteristics of the Items

The simulated test had 66 items, a common length for "real world" tests. Item discrimination was set so that the first 33 items had an "a" value of 0.80, and the second 33

items had an "a" value of 1.20. These values represent the upper and lower bounds of a range within which most items in real world tests fall (Hambleton & Swaminathan, 1985, p. 36).

It has been found that the "b" values for most real world tests range from -2.00 to 2.00 (Hambleton & Swaminathan, 1985; p. 36), so the "b" values for the items answered by the RG were generated as follows: the first of the 66 items on the test was given a value of -2.00, with each successive item having a "b" value 0.125 greater than its predecessor, until item 33 had a "b" value of 2.00. The process was then repeated starting with item 34 and proceeding through to item 66.

The values of "b" for the 66 items on the test answered by the FG had "b" values generated in the same way, except that four out of the 66 items had a "b" value higher for the FG than the same item for the RG, thereby creating uniform dif that favored the RG over the FG. There were four levels of difference in "b" between the FG and RG of .25, .50, .75, and 1.00, respectively. The items selected to be made to exhibit dif were those two items on each 33-item half of the test which had "b" levels of -0.75 and 0.75 for the RG examinees, for a total of four items exhibiting dif per 66-item test. The pseudo-guessing ("c") parameter was held constant at .15

Power Analysis

The degree of dif, expressed above as differences in level of "b" between RG and FG, can also be expressed in two other ways: as a difference in p-values between RG and FG, or in terms of the concept of "effect size" according to the classification proposed by Cohen (1992).

In Cohen's typology, effect size can be derived from a difference in decimal proportions (like the difference in p-values) by finding the difference between the arcsine transformations of those proportions.

Since effect size in this study was not initially set as a proportion, but was instead described as a difference in level of "b" between RG and FG, the differences in "b" were translated into decimal proportions by following the procedure outlined in Crocker and Algina (1986).

The first step consisted of determining the biserial correlation (ρ_g) which corresponded to the "a" values chosen for the study, according to the formula:

$$A_G = \frac{\rho_G}{\sqrt{1-\rho^2{}_G}} \tag{12}$$

The values for ρ_g derived from this equation, using "a" values of 0.80 and 1.20, were then inserted into formula 12 in order to derive the p-value that corresponded to the appropriate level of "b":

$$b_g = \frac{-\Phi^{-1}(p_g)}{\rho_g} \tag{13}$$

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

By inserting the appropriate value of "b", which was set by the researcher, and the value of ρ_g derived from equation 11, a z-score value was found. The area to the left of that z-score

was then determined, this area being the equivalent of the p-value of classical test theory. The p-values were determined in this way for each level of "b" in the null condition (-0.75 and 0.75) as well as each of the values of "b" in the biased conditions. These p-values were derived for the "b"s under each of the two levels of "a", giving a total of twenty p-values, including the p-values for the null condition. These p-values were corrected in order to take into account the inflation caused by the presence of a pseudo-guessing parameter of 0.15, according to the following formula:

$$p_c = p + (1-p)c \tag{14}$$

where p_c refers to the p-value corrected for guessing.

The arcsine transformations of these corrected p-values were then used in the formula given in Cohen (1992), for the determination of the effect size of a difference between two decimal proportions:

$$h = \text{arcsine}A - \text{arcsine}B \tag{15}$$

where h is effect size, A is the item difficulty expressed as a p-value for the RG (i.e., those p-values derived from the null "b" levels of -0.75 and 0.75 respectively), and B is the item difficulty for the FG (i.e., those p-values derived from the biased "b" levels).

The effect sizes, "h", in this study ranged from .0519 to .3291. In Cohen's typology, an effect size of .20 is considered "small", while an effect size of .50 is "medium". Therefore, the values of "h" in this study can be considered to range from small to somewhat less than medium. The differences in p-values between RG and FG before transformation to "h" ranged from .048 to .250. As can be seen from Table 2 below, which shows the p-values and h-values derived from the process of converting differences in "b" to "h", for any given combination of "a" and difference in "b", both p-value difference and "h" were always higher when "b" was low ("b" = -0.75), than when "b" was high ("b" = 0.75).

Data Analysis

The MH indices were estimated using the program developed by Ackerman (1986), and the LR indices by the program developed by Spray (1991). Spray (1991) calculates uniform and non-uniform dif separately and therefore employs two separate chi-square indices, each with one degree of freedom. Levels of significance were set at .01 and .05 for both procedures. However, only the results at the .01 level of significance were analyzed.

A total of one hundred replications were run at each of the five sample sizes. On any given run, all four of the items which exhibit dif had only one of the four possible effect sizes. This means that four successive runs of one hundred replications at each sample size needed to be performed in order that all four levels of effect size and all five levels of sample size could be examined together. The effect of item discrimination was controlled for by the fact that two levels of "a" (0.80 and 1.20) were present in every 66-item test.

Table 2. Relationship between level of "b", level of "a", difference in "b", p-value, p-value corrected for guessing, p-value difference RG-FG, and "h".

"b" value diff.	"b"	"a"	ρ	p-value	p-value corr.	p-diff RG-FG	" h"
0.00 (null)	-0.75	0.80	0.625	0.680	0.728		
	-0.75	1.20	0.768	0.718	0.760		
	0.75	0.80	0.625	0.332	0.422		
	0.75	1.20	0.768	0.282	0.390		
0.25	-0.50	0.80	0.625	0.623	0.680	0.048	.0676
	-0.50	1.20	0.768	0.648	0.701	0.059	.0865
	1.00	0.80	0.625	0.264	0.374	0.048	.0525
	1.00	1.20	0.768	0.223	0.340	0.050	.0537
0.50	-0.25	0.80	0.625	0.561	0.627	0.101	.1377
	-0.25	1.20	0.768	0.575	0.639	0.121	.1701
	1.25	0.80	0.625	0.217	0.334	0.088	.0953
	1.25	1.20	0.768	0.169	0.294	0.096	.1022
0.75	0.00	0.80	0.625	0.500	0.575	0.153	.2028
	0.00	1.20	0.768	0.500	0.575	0.185	.2507
	1.50	0.80	0.625	0.175	0.299	0.123	.1322
	1.50	1.20	0.768	0.125	0.256	0.134	.1417
1.00	0.25	0.80	0.625	0.438	0.522	0.206	.2662
	0.25	1.20	0.768	0.423	0.510	0.250	.3281
	1.75	0.80	0.625	0.137	0.266	0.156	.1666
	1.75	1.20	0.768	0.091	0.227	0.163	.1716

Regression analyses were conducted, using the MHCHISQ, LRU, and Δ_{MH} as outcome variables. Level of "b" for the RG, level of "a", difference in "b", effect size (h), difference in p-value between RG and FG, and sample size, were the predictor variables. Percentage dif detection rates were reported for each of the indices studied. False positive rates for the MH and LR were also reported.

CHAPTER 4

RESULTS AND DISCUSSION

The results of the study are reported and discussed in this chapter. The chapter is divided into two sections. The first section contains a summary and discussion of the results of the regression analysis predicting the value of Δ_{MH}, and the dif detection rates of the MH and LRU. Since the dif generated in the study was uniform, no attempt was made to construct a regression model predicting LRN dif detection rates, since the LRN is not designed to be able to detect that type of dif. The second section is devoted to summarizing and discussing the effects of the studied variables on dif detection rates for the MH and LRU. In addition, the second section includes a brief discussion of the false positive rates of each of the detection procedures.

Regression Analysis

Regression analyses were performed using as outcome variables MH dif detection rate, LRU dif detection rate, and Δ_{MH} value. Detection rate was the percentage of times the items with uniform dif were detected over the 100 replications performed. Δ_{MH} value was the mean Δ_{MH} over 100 replications. The predictor variables were effect size (EFF), p-value difference between RG and FG (PD), the difference in level of "b" between RG and FG (BD), sample size (SS), level of "b" for the RG (RGB), and level of "a" (A).

Only results at significance level .01 were analyzed. This was done because at significance level .05, both the MH and LRU achieved a very high level of success at the larger sample sizes (Tables 11 to 15, Appendix A), leading to a ceiling effect that suppressed the variability of the distributions of the outcome variables. This effect was less pronounced at significance level .01, permitting a more valid weight to be attached to effect of sample size on the outcome variables.

Correlation coefficients for the predictor variables are shown in Table 3. The dif detection rates of the MH and LRU were highly correlated (r=.990). The detection rates of the MH and LRU were correlated quite highly with the value of the Δ_{MH} ($r_{MH \cdot \Delta MH}$= -.657, $r_{LR \cdot \Delta MH}$ = -.690). The correlations were negative because dif was generated to favor the RG. Table 3 also shows that the Δ_{MH} was very highly correlated with the three measures of degree of dif ($r_{\Delta MH \cdot EFF}$= -.986, $r_{\Delta MH \cdot PD}$ = -.968, $r_{\Delta MH \cdot BD}$ = -.803).

Table 3. Correlation matrix for the four outcome variables and the seven predictor variables.

Pearson Product-Moment Correlation (r)

	MH	LRU	Δ_{MH}	EFF	PD	BD	SS	RGB	A
MH	1.000								
LRU	**.990**	1.000							
Δ_{MH}	**-.657**	**-.690**	1.000						
EFF	**.665**	**.700**	**-.986**	1.000					
PD	**.688**	**.727**	**-.968**	**.974**	1.000				
BD	**.654**	**.703**	**-.803**	**.813**	**.916**	1.000			
SS	**.571**	**.513**	.015	.000	.000	.000	1.000		
RGB	-.202	-.208	**.435**	**-.483**	-.286	.000	.000	1.000	
A	.132	.120	-.281	.149	.146	.000	.000	.000	1.000

bold = $p < .01$

This result was expected, given that Δ_{MH} provides a measure of both the direction and degree of dif (Holland & Thayer, 1986), and is therefore very similar to the variables PD, EFF, and BD. Since the correlations of PD, EFF, and BD with the other dif detection indices were very similar to those of Δ_{MH} with those same indices, this conclusion seemed justified. RGB was significantly inversely correlated with Δ_{MH} and EFF, ($r_{RGB \cdot \Delta MH} = .435$, $r_{RGB \cdot EFF} = -.483$), while A was not significantly correlated with any other variable. Sample size was correlated with only the outcome variables due to the design of the study, but its correlation was relatively low, given the ceiling effect mentioned earlier.

The six predictor variables produced 15 two-way interactions. In order to determine if interactions between the predictor variables had an effect on predictive success, a variable-by-variable analysis was performed according to the procedure described in Darlington (1990). The first step of the procedure consisted of performing a single test for every interaction that involved each of the predictor variables, and then correcting for the resulting alpha-inflation. The tests performed compared the changes in R^2 between the model without interaction terms and with interaction terms. Since there were six regressors, six comparisons were performed, and were corrected using the Bonferroni method.

As can be seen from Table 4, none of the models that incorporated interactions explained significantly more variance in MH detection rates than did the full six-variable model without interactions. Therefore, the regression model for the prediction of MH dif detection rate was based on the all-possible subsets analysis that used only the six predictor variables without interactions. However, interactions involving PD and EFF were significant in predicting dif detection rates for the LRU, so the interactions that involved these variables were incorporated

into the follow-up analysis. Interactions involving all the predictor variables except sample size had significant effects on the prediction of mean ΔMH value. Since the Δ_{MH} value will approach the mean in every case, the lack of relationship with sample size is is not surprising.

Table 4. R^2 values for the variable-by-variable tests of all possible two-way interactions predicting MH and LR detection rates, and mean Δ_{MH} values.

Model	MH	LR	Δ_{MH}
PD + EFF + BD + SS + RGB + A	.8199	.8231	.9920
PD + EFF + BD + SS + RGB + A+ PD*EFF + PD*BD + PD*SS + PD*RGB + PD*A	.8523	**.8746**	**.9960**
PD + EFF + BD + SS + RGB + A + EFF*PD+EFF*BD+EFF*SS+EFF*RGB+EFF*A	.8549	**.8784**	**.9961**
PD + EFF + BD + SS + RGB + A + BD*PD + BD*EFF + BD*SS + BD*RGB + BD*A	.8515	.8683	**.9960**
PD + EFF + BD + SS + RGB + A + SS*PD + SS*EFF + SS*BD + SS*RGB + SS*A	.8437	.8468	.9926
PD + EFF + BD + SS + RGB + A + RGB*PD+RGB*EFF+RGB*BD+RGB*SS+RGB*A	.8570	.8715	**.9963**
PD + EFF + BD + SS + RGB + A + A*PD + A*EFF + A*BD + A*SS + A*RGB	.8226	.8254	**.9964**

bold = $p < .0017$

MH Detection Rate

The full regression model without interactions was significant at $p < .001$, and accounted for 81.99% of the variance in MH dif detection rate. However, three components of the model, EFF, PD, and BD, were simply different measures of the same construct, the amount of induced

dif in the items of the test. Since all three variables were highly correlated with one another ($r > .80$), much of their individual contribution to the overall model was overlapping. Conceptually, it appeared more useful to examine the models that contained each of these variables separately from one another, but which also included the other variables of interest, SS, RGB, and A. These were the four-variable models in the all-possible subsets analysis.

The model which included BD, SS, RGB, and A explained 81.23% of the variance in MH dif detection, nearly as much as the overall model. The model with PD instead of BD explained 79.99% of the variance, while the model with EFF instead of BD or PD explained 78.73% of the variance. However, the four-variable models included RGB and A as variables. Due to the design of the study, BD was not correlated at all with either RGB or A, while both PD and EFF were negatively correlated with RGB ($p < .01$ and $p < .05$ respectively), and were positively correlated with A. The correlations among the variables PD and EFF with RGB and A would have tended to suppress the resulting R^2 (Darlington, 1992), due to the phenomenon of collinearity. For this reason, the apparently superior performance of the four-variable model incorporating BD must be interpreted with caution.

The most parsimonious models appeared to be the two-variable models PD with SS, EFF with SS, and BD with SS (Table 5). Each of these models explained more than 75% of the variance in MH dif detection rate, with the best of the models, PD with SS, explaining almost 80%, only 2% less than the overall six-variable model. It was determined that adding another predictor (BD) to the model PD with SS did not significantly improve the amount of explained variance ($F(2,76) = 1.38$, $p > .05$). For purposes of illustrating the absence of any effect of interactions on MH dif detection rates, it should be pointed out that adding PD*SS, the interaction

of PD with SS, to the model increased the explained variance to 80.08%, an improvement of only 0.09%.

Table 5. Summary of the best one-, two-, and three variable models explaining MH DIF detection rates, together with the overall six-variable model.

Model	R^2
PD	.4730
EFF	.4425
BD	.4282
PD + SS	.7988
EFF + SS	.7683
BD + SS	.7539
PD + SS + BD	.8024
PD + SS + P*S	.8008
PD + EFF + BD + SS + RGB + A	.8199

Other than the three models PD with SS, EFF with SS, and BD with SS, no two-variable model explained more than 48% of the variance in the outcome variable. The best one-variable model, PD alone, accounted for only 47.30% of the variance in the outcome variable. It was confirmed that PD with SS accounted for significantly more variance ($F(2,77)=125.27$, $p<.01$) than PD alone. Given that the MH is known to be strongly affected by fluctuations in sample size (Brown, 1992; Ibrahim, 1992; Pang et al., 1994; Tian et al., 1994a, 1994b), and also by the degree of dif (Clauser et al., 1991a; Mazor et al., 1991), the two-variable model described above

was, therefore, not only the most parsimonious model, it was also conceptually quite plausible.

It was possible, moreover, that the predictive power of all three of the best two-variable models mentioned was suppressed by a ceiling effect involving sample size (SS). The MH detected the dif items almost all the time at the higher sample sizes, especially when the amount of dif was large (see discussion below and Tables 11 to 15 in Appendix A). Since ceiling effects suppress the variability of the dependent variable, the R^2s of the models with SS would also have been suppressed (Darlington, 1992). It appeared that any two-variable model using SS as one of its variables would actually have explained more variance if it had been applied only to the smaller sample sizes used in the study.

Of the three two-variable models described above, the model PD with SS explained the most variance in MH dif detection rate. In addition, PD was more highly correlated with MH detection rate (r = .6878) than were either EFF (r = .6652) or BD (r = .6543). It appeared, therefore, that PD (p-value difference between RG and FG) was a better predictor of MH dif detection rate than either EFF (effect size) or BD (difference in level of "b" between RG and FG). Since the inclusion of PD*SS would not involve the use of another predictor variable in the model, it was decided to include PD*SS in the final model, even though it added little more explained variance.

The complete model using PD with SS and PD*SS is as follows:

$$MH = -38.032 + 495.269(PD) + .098(SS) - 0.115(PD*SS)$$

The corresponding standard errors of the b-weights were 63.944, 0.0178, and 0.013 ,

respectively. The standard error of the estimate was 17.25.

In order for a regression analysis to be performed and results to be interpreted validly, the assumptions of normality, linearity, equality of variances, and independence must not be violated. For this model, the residual plots indicated that the distributions generally fit these assumptions. However, the ceiling effect for sample size was present, and the residual plots indicated an increase in the magnitude of the residuals as sample size increased. However, Cook's D was consistently low ($D < 0.56$), indicating that no particular data point exerted an overly strong influence on the distribution as a whole. The b-weights of the model needed to be interpreted with some caution due to scaling factors. The b-weight for PD was very large since the actual p-value differences between RG and FG were quite small, ranging from .048 to .250. Since PD alone accounted for 47.3% of the variance, its b-weight was inflated. The b-weight for sample size, however, was artificially deflated since sample size in the study ranged from 100 to 800.

The model can be used to predict the rate of dif detection given certain conditions of sample size and p-value difference, assuming that the test is in fact a 66-item test. For example, a test developer with 600 examinees may be interested to know what percentage of dif would be detected if the dif is low to moderate (PD = .060) and the interaction between PD and SS is small.. In this case, the regression model would be as follows (with no interaction term):

$$MH = -38.032 + 495.269(.060) + .084(600)$$

The resulting detection rate of 42% can be used as the basis for confidence interval estimation. However, the large standard error of the estimate (17.22) makes useful confidence

interval estimation problematic. The model can solve for the effect of sample size on dif detection, given a specific level of dif, for example, PD = .100, and a specified degree of success, for example, 80%. In this case, the regression model would be as follows:

$$80 = -38.032 + 495.269(.100) + .084(SS)$$

Solving for the above equation shows that a sample size of 815, plus or minus the appropriate confidence interval, would be needed if the test administrator requires a success rate of 80% dif detection.

LRU Detection Rate

The LRU detection rate model had the same six predictor variables as the MH detection model described above. The overall six-variable model without interaction terms was significant ($p < .001$), and accounted for 82.31% of the variance in LRU dif detection rate. This was slightly more than the overall MH detection rate model. As with the MH, the best two variable model was PD with SS, and it accounted for 79.16% of the variance in LRU dif detection rate, almost as much variance as the overall model without interactions. The best one-variable model was BD*SS, the interaction of BD with SS. However, it accounted for only 58.26% of the variance in LRU detection rate. PD alone accounted for 52.86% of the variance in the outcome variable. It was determined that PD with SS accounted for significantly more variance in LRU dif detection rate than did PD alone ($F(2,77) = 97.41$, $p < .01$), and the absolute improvement in explained variance was also substantial (> 26%).

An inspection of Table 4 shows that the models that included all six predictor models with the interactions involving PD explained significantly more variance in LRU dif detection rate than did the six-variable model alone. Since the LRU was more highly correlated with PD than it was with any other predictor variable (r_{LRU*PD} = .727), this finding was not unexpected.

Table 6. Summary of the best one-, two-, and three variable models explaining LRU DIF detection rates, together with the overall six-variable model.

Model	R^2
PD	.5286
PD + SS	.7916
PD + SS + PD*SS	.8015
PD + EFF + BD + SS + RGB + A	.8231

Table 6 shows that adding another variable (PD*SS) to the model PD with SS improved the amount of explained variance from 79.16% to 80.15%, a difference of less than 1%. It was confirmed that this increase was statistically significant ($F(3,76) = 3.81$, $p < .05$), so the three-variable model was employed. As with the MH dif detection model, the model PD with SS and PD*SS had the advantage of conceptual plausibility as well as parsimony. Several studies have found that the LRU is susceptible to sample size fluctuations (Brown, 1992; Ibrahim, 1992; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Pang et al., 1994; Tian et al., 1994a, 1994b), as well as being more effective at dif detection when dif is relatively large (Pang et al., 1994; Tian et al., 1994a, 1994b). As with the MH model, residual scatterplots and Cook's

D were examined and indicated that the assumptions of linearity, homescedasticity, independence, and equality of variances were met.

The complete model with b-weights is as follows:

$$LRU = -34.52 + 548.64(PD) + .1011(SS) - .2395(PD*SS)$$

The standard errors of the b-weights were 60.48, .0168, and .1229, respectively, and the standard error of the estimate was 16.31. As with the MH dif detection model, the b-weights must be interpreted in light of their scale-boundedness. The model can be used to generate predictions as above. It must also be pointed out that the LRU had a somewhat higher rate of dif detection than the MH under almost every condition of the study. The residual plots indicated that the sample size ceiling effect mentioned earlier was again present. It is possible that the variance explained by PD with SS and PD*SS for the LRU was slightly suppressed compared to the MH, resulting in the slightly smaller amount of variance explained.

Δ_{MH}

The full model without interactions for predicting the value of Δ_{MH} had the same six predictor values as the three models above. This model was able to account for 99.20% of the variance in the value of Δ_{MH}, and was significant ($F(6,73)=1500.55$, $p<.001$). However, the best of the one-variable models from the all-subsets analysis with interactions, EFF alone, was able to account for 97.18% of the variance. Adding EFF*A, the interaction of effect size with item discrimination, to the model improved the amount of variance accounted for to 99.33%, more than

the entire six-variable model without interactions. It was determined that adding this second variable to the one-variable model did significantly improve the amount of variance explained ($F(3,76) = 248.28$, $p < .01$). However, the improvement was small in practical terms, slightly more than 2%. For reasons of economy, therefore, the one-variable model was preferred. The complete model with b-weights follows:

$$\Delta_{MH} = -0.015489 - 10.4102 \text{ (EFF)}$$

The standard error of the b-weight was .201. The standard error of the estimate was 0.13827. Inspection of the residual scatterplots indicated that some inequality of variances existed at the upper end of the distribution, but it did not appear to be sufficiently large to invalidate the use of regression techniques.

From the very high correlations of Δ_{MH} with EFF and PD, and to a lesser extent with BD (Table 3), it seemed appropriate to conclude that, in some sense, ΔMH is virtually the same measure as EFF and PD. PD alone and EFF alone each explained a very high amount of the variance in Δ_{MH} (97.175% and 93.76%, respectively), while BD explained only 64.52%. Both PD and EFF are measures of effect size derived from BD, but which incorporate some of the effects of differences in "b" and "a". Since this modification of BD into PD and EFF results in a single-variable model capable of explaining more than 90% of the variance in ΔMH, it appears that PD and EFF are better measures of effect size than BD.

DIF Detection Rates

This section includes a discussion of the effects of the various predictors on the dif detection rates of the MH and LRU.

P-value difference and sample size

As shown in Table 7, the MH was able to detect dif items at sample size 800 and .01 level of significance 80% or more of the time when the p-value difference was .088 (i.e., in the small to medium range of the p-value differences in the study) . At sample size 600 and significance level .01, the MH was still able to detect dif items more than 80% of the time or more when p-value difference was .096. When sample size was 400 and significance level was .01, detection rates exceeded 80% only when p-value differences were .12 or larger. At sample size 200 and .01 level of significance, it took a p-value difference of .12 or more (about the middle range for this

Table 7. Level of difference in p-value required for the MH and LRU to achieve a DIF detection success rate of 80 per cent or greater.

	Significance level = .05		Significance level = .01	
Sample Size	MH	LRU	MH	LRU
800	.059	.059	.088	.088
600	.088	.088	.096	.096
400	.096	.096	.121	.101
200	.121	.121	.185	.163
100	.206	.121	.250	.206

study) for the MH to detect dif more than 80% of the time. At .01 level of significance, only p-value differences greater than .18 produced a detection rate greater than 80%. At sample size 100, at both levels of significance, only differences at the upper range of the values used in this study (.20 to .25) produced success rates higher than 80%.

The LRU performed similarly to the MH, detecting dif all or almost all the time (80% or more) at sample size 800 whenever the difference in p-value was .088 (.01 level of significance) or .059 (.05 level of signficance). At sample size 600 and significance level .01, a p-value difference of .096 or greater continued to produce detection rates in excess of 80%. At sample size 400, significance level .01, an 80% or greater success rate was attained whenever p-value difference was .096 or higher. Even at sample size 200, when significance level was .01, a p-value difference of only .163 was sufficient to produce a success rate of 80%. At sample size 100 and significance level .05, a p-value difference of more than .20 was needed to achieve a detection rate higher than 80%.

In general, both the MH and LRU performed well, even at small sample sizes, if the p-value difference was above the midrange of values used in the study (about .12). These values, while small in absolute terms, are comparable to the p-values discussed in Mazor et al. (1991) that were detected by the MH when the RG and FG ability distribution were equal. It appears that the findings of the study are approximately comparable to those of Mazor et al.

Swaminathan and Rogers (1990) found that the MH and LR procedures were about 75% likely to detect uniform dif at sample size 250, and 100% at sample size 500. These figures are somewhat less than those found in this study. Table 8 shows that at sample size 600, the MH and LR were able to detect dif items 78.13% and 79.94% of the time, respectively. In their 1993

study, Swaminathan and Rogers had results that indicated that the MH performed marginally better than the LR at dif detection, although this difference with the current study may have been the result of the different computer programs used to compute the statistical indices.

Table 8 shows that the LRU detected dif items more often at every sample size, with the difference in dif detection rates becoming even more pronounced as sample size dropped. The difference in dif detection rates of a little more than 1% at sample size 800 increased to 4% at sample size 400, and at sample size 100, the difference had grown to almost 10%. This finding contradicts Rogers and Swaminathan (1993), who found the MH marginally better than the LR, but supports Pang et al. (1994), and Tian et al. (1994a, 1994b).

Table 8. Mean % DIF detection, MH and LRU, across sample sizes, significance level .01.

DIF Index	Sample Size				
	800	600	400	200	100
MH	82.25	78.13	67.31	44.00	21.13
LRU	83.44	79.94	71.25	51.68	30.81

Item Difficulty and Item Discrimination

The MH and the LRU detected dif items more often whenever "a" was large ("a" = 1.20), a finding similar to that of Mazor et al., who found that low discrimination items tended to be missed even at large sample sizes. The MH and LRU also tended to detect low difficulty items more often than high difficulty items. The reason for this tendency can be understood by

returning to the concept of "effect size", and relating it to "a" and "b".

The effect size column (i.e., "h") in Tables 11 to 15 (Appendix A) shows that across every sample size and every level of difference in "b", the effect size was directly related to the detection rates of both the MH and LRU. The low difficulty, high discrimination item consistently had the highest effect sizes of any of the items under all conditions. The high difficulty, low discrimination item had the lowest effect sizes of any item under all conditions. It was also found that the low difficulty items, considered together, had larger effect sizes than the high difficulty items, and the high discrimination items, considered together, had larger effect sizes than the low discrimination items. Since the effect sizes were computed from p-values, they are related to the areas between ICCs (Raju, 1988), i.e., the larger the effect size, the larger the area between the ICCs for the RG and FG. It is clear from the results of this study that highly discriminating items of low difficulty must have the largest areas between ICCs for the RG and FG, and this is why both the MH and the LRU have the greatest degree of success in detecting this type of item. Similarly, items of high difficulty are items at the upper end of the ICC, and have a smaller area between the ICCs, especially when the level of "a" is low. This type of item was clearly the most difficult to detect for both the MH and the LRU.

Table 9 shows clearly that both the MH and LR were more able to detect dif when item difficulty was low than when it was high. However, both indices were better able to detect dif items when item discrimination was high. The finding that low discrimination adversely affected the MH was in agreement with Mazor et al. (1991), as was the finding that the items most likely to be missed were those of high difficulty, low discrimination, and small differences in "b".

Table 9. Mean % DIF detection, MH and LRU, at different levels of "b" for RG and levels of "a", significance level .01

DIF Index	Level of "b" for RG		Level of "a"	
	-0.75	0.75	0.80	1.20
MH	66.18	50.95	53.58	63.55
LRU	73.35	57.30	59.15	67.7

However, Mazor et al. also found that at sample size of 500, more than half the dif items were missed. This result was not duplicated in this study, where it was found that at sample size of 400, the MH detected the dif items 80% of the time or more at significance level .05 whenever the difference in p exceeded .10. This finding is even more in disagreement, however, since Mazor et al. used levels of difference in "b" that were greater than those employed in this study. However, they also used "a" values of .25, 60, .90, 1.25, some of which were considerably smaller than those used in this study. Since the MH is less effective when "a" is low, this may have affected the dif detection rate adversely in the Mazor et al. study.

Clauser, Mazor, and Hambleton (1991a) also found that high discrimination led to substantial improvements in the ability of the MH to detect dif items, while high levels of "b" tended to depress the success rate of the MH. Both of these findings are compatible with those of this study.

False-Positive Rates

The results of the false-positive analysis are presented in Table 10. At sample size 800 and difference in "b" of 1.00, the MH and the LR exceeded the expected false-positive rate. Each index approached the expected false-positive rate as difference in "b" dropped. The same results were obtained at sample size 600, although both the MH and LRU approached their expected false-positive rate when difference in "b" was as small as .50. At sample size 800 and 600, it became apparent that the MH tended to have lower false-positive rates than the LRU. This relationship held as sample size declined. This result bears out a finding of Tian et al. (1994a, 1994b), and Pang et al. (1994), that while the LRU tends to be more powerful than the MH, it has a higher false-positive rate. In fact, there was never a sample size or level of difference in "b" at which the false-positive rate of the MH exceeded that of the LRU.

It appeared from the false-positive analysis that the moderately superior dif detection performance of the LRU over the MH had a price: a tendency to flag non-dif items as exhibiting dif. At small sample sizes, the discrepancy between the false-positive rates of the two procedures was fairly substantial. For example, at sample size 100 and effect size of .25, the MH had a false-positive rate of 2.97, well under the expected rate, while the LRU had a false-positive rate of 5.35, about the expected rate. It appeared that the decline in power of the MH as sample size decreased caused the MH to become much less susceptible to false-positive identifications. Indeed, at smaller sample sizes and effect sizes, the MH often had a lower false-positive rate than expected.

TABLE 10. Percentage False Positive Identifications for the MH, and LRU by Sample Size and Difference in "b".

Sample Size	Diff. in "b"	Significance level .05		Significance level .01	
		MH	LRU	MH	LRU
800	1.00	8.58	9.24	2.26	2.5
	0.75	6.48	7.44	1.65	1.94
	0.50	5.53	6.31	1.26	1.53
	0.25	4.6	5.61	.90	1.1
600	1.00	7.21	8.35	2.02	2.26
	0.75	6.02	7.11	1.45	1.77
	. 0.50	5.11	5.81	0.89	1.13
	0.25	4.74	5.48	0.77	0.97
400	1.00	5.10	6.74	1.16	1.55
	0.75	4.58	5.95	0.76	1.26
	0.50	4.45	5.81	0.82	1.15
	0.25	4.29	5.53	0.94	1.15
200	1.00	4.11	5.73	0.58	1.08
	0.75	3.80	5.06	0.79	1.15
	0.50	3.55	5.29	0.53	1.02
	0.25	3.71	4.95	0.79	1.10
100	1.00	2.90	5.69	0.45	0.90
	0.75	3.15	5.48	0.47	1.08
	0.50	2.65	4.87	0.47	0.95
	0.25	2.97	5.35	0.47	1.02

CHAPTER 5

CONCLUSIONS

Summary of the Findings

The results of the study clearly indicated that the MH and LRU procedures were highly reliable detectors of dif at sample sizes of 600 and up, even when effect size (considered as either "h" or as differences in p-values) was relatively small. When sample size was 400 and effect size was relatively large, the success rate of both procedures was high. If the less stringent significance level of .05 is used, then dif items were detected successfully by both the MH and LRU more than 80 per cent of the time whenever the difference in level of "b" was as large or larger than 0.50 and sample size was as small as 400. Even at the sample sizes of 200 and 100, if effect size was larger, the detection rates of both procedures were high, approaching or exceeding 70 per cent. However, the MH appeared to be less successful than the LRU at dif detection when sample sizes were very small and effect size was also small. For example, when sample size was 200 and effect size was large, the difference in detection rates between the MH and the LRU was small. When sample size was 100, the difference in detection rates between the procedures was proportionally much larger. It appeared that the LRU was somewhat more resistant to sample size effects than the MH. These findings were similar to those of Pang et al. (1994), who concluded from their research that the MH and LR procedures can be expected to detect dif reliably at sample sizes of 500 and larger.

An inspection of Tables 11 to 15 in Appendix A yielded the finding that while the MH and LRU were often equal in their dif detection rates, whenever a difference existed it nearly always favored the LRU over the MH. With only a single exception, under all sample sizes and effect sizes, over every combination of item difficulty and item discrimination, at both levels of significance, the LRU flagged the dif items more often than did the MH. While the difference between the two procedures' success rates was never extremely large, it appeared that the LRU was moderately more powerful than the MH at detection of uniform dif under the conditions in the study, especially at the smaller sample sizes, where the difference in their success rates appeared to increase.

However, as in Tian et al. (1994b) and Pang et al. (1994), the LRU had a higher false positive rate than the MH as well. It was concluded from the false-positive analysis that the power of the LRU as compared to the MH was directly related to its tendency to incorrectly flag non-dif items. If a test developer who wishes to make a choice between the LRU and the MH is more desirous of ensuring that all dif items are excluded from a test, then the LRU would appear to be the superior procedure. However, this superiority is more clearly manifest at smaller sample sizes and at smaller effect sizes, where neither procedure may be considered sufficiently reliable to be employed, and where the degree of dif may be so small that it would have relatively little effect on test outcomes in any case.

The regression analyses showed that an acceptable model can be derived for both the MH and LRU using only two variables, difference in p-value between RG and FG, together with sample size. For the Δ_{MH}, effect size was the only predictor variable needed to construct the regression model. The regression analyses demonstrated that the traditional method of measuring

effect size, difference in level of "b", was inferior to each of the two alternatives employed in this study, especially to the difference between p-values for RG and FG.

Limitations of the Study

The greatest limitation of this study was also its greatest strength, the use of simulated data. Whiie the use of simulated data made conclusions easier to draw and facilitated the formulation of fairly parsimonious regression models for the MH and LRU procedures, the possibility remains that other variables of interest were not properly taken into account. Furthermore, the specific values attributed to the variables may not be truly reflective of real data. Further replications of this study would be desirable, perhaps focusing in more detail on the effect of item discrimination and item difficulty on dif detection procedures, as well as engaging in a more in-depth analysis of the virtues of the various methods used for measuring effect size and their relationship to ΔMH. In particular, the use of "h" as a measure of effect size may not be possible in data derived from real world test administrations. In order to compute "h" accurately, "a" and "b" values must be derived for the examinees. If this cannot be done, then $\Delta_{MH,}$ which is very highly correlated with "h", could perhaps be employed instead. Research into this question would be useful. It might also be of interest to examine different lengths of test and different levels of dif, since these variables would be of prime importance if the results of this study are to be validly generalized to the "real world".

In general, the MH and the LRU were quite successful at detecting dif at sample sizes of 800 and 600. So successful were they that a ceiling effect was created, suppressing the variability

of the outcome variables in relation to sample size. Given this result, a more optimal model might have been constructed if logistic regression methods had been employed, rather than the linear regression methods that were used in the study. In addition, restricting sample sizes to less than 500 might have produced a more adequate and useful model.

CHAPTER 6

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.

Ackerman, T.A. & Evans, J. (1992). An investigation of the relationship between reliability, power, and the Type I error rate of the Mantel-Haenszel and Simultaneous Item Bias detection procedures. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA, April, 1992.

Baghi, H. & Ferrara, S. (1989). A comparison of IRT, delta plot, and Mantel-Haenszel techniques for detecting differential item functioning across subpopulations in the Maryland Test of Citizenship Skills. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, March, 1989.

Baghi, H., & Ferrara, S.F. (1990). Detecting differential item functioning using IRT and Mantel-Haenszel techniques: Implementing procedures and comparing results. Paper presented at the annual meeting of the Eastern Educational Research Association, Clearwater, FL, February, 1990.

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

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

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

Clauser, B., 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, 1991.

Clauser, B., Mazor, K.M., & Hambelton, 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.

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

Gutierrez, J. (1988). Characteristics of the Mantel-Haenszel delta under different conditions of the null hypothesis: A Monte Carlo study. Unpublished M.A. thesis, University of Ottawa, Ontario, Canada.

Hambleton, R.K., & Jones, R.W. (1992). Comparison of empirical and judgmental methods for detecting differential item functioning. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA, April, 1992.

Hambleton, R.K., & Rogers, H. J. (1988). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2 (4), 313-34.

Hambleton, R.K., Rogers, H.J., & Arrasmith, D. (1988). Identifying potentially biased test items: A comparison of the Mantel-Haenszel statistic and several item repsonse theory methods. Paper presented at the annual meeting of the American Educational Research Association, San Fransisco, CA, April, 1988.

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

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

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

Kromrey, J.D., & Parshall, C.G. (1991). Screening items for bias: An empirical comparison of the performance of three indices in small samples of examinees. Paper presented at the annual meeting of the Florida Educational Research Association, Clearwater, FL, November, 1991.

Linacre, J.M., & Wright, B.D. (1987). Item bias: Mantel-Haenszel and the Rasch model. Finnish Association of Mathematics and Science Education Research, Memorandum 39.

Mazor, K.M., Clauser, B., & 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, 1992.

Mellenburgh, G.J. (1982). Contingency tables models for assessing item bias. Journal of Educational Measurement, 18, 229-248.

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

Pang, X.L., & Boss, M.W. (1993). The effects of sample size, item difficulty, and item discrimination on logisitic regression item bias indices. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, Georgia, April, 1994..

Pang, X.L., Tian, F., & Boss, M.W. (1994). Performance of the Mantel-Haenszel and logistic regression procedures over replications using real data. Paper presented at the annual meeting of the American Educational Research Association, April 1994.

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

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

Ryan, K.E. (1990). The performance of the Mantel-Haenszel procedure. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA, April, 1990.

Shephard, L., Camilli, G., & Williams, D.M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22(2), 77-105.

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

Sudweeks, R.R., & Tolman, R.R. (1990). The use of empirical versus subjective procedures for identifying science test items which function differentially for females and males.

Paper presented at the annual meeting of the National Association for Research in Science Teaching, Atlanta, GA, April, 1990.

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

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

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

Tian, F., Pang, X.L., & Boss, M.W. (1994a). The consistency of the Mantel-Haenszel and logistic regression DIF identification procedures across sample size and overreplications. Paper presented at the annual meeting of the American Educational Research Association, April, 1994.

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. Paper presented at the annual meeting of the National Council on Measurement in Education, April, 1994.

APPENDIX A

Table 11. MH and LRU DIF detection rates, and Δ_{MH} value, sample size 800.

Null "b"[1]	"a"[2]	Diff. "b"[3]	Eff. Size[4]	Diff. p[5]	Signif. level .01[6]		Signif. level .05[7]		Δ_{MH}[10]
					MH[8]	LRU[9]	MH[8]	LRU[9]	
-.75	0.80	1.00	.2659	.206	100	100	100	100	2.50
		0.75	.2028	.153	100	100	100	100	1.97
		0.50	.1377	.101	97	100	100	100	1.32
		0.25	.0676	.048	27	35	58	64	0.62
	1.20	1.00	.3291	.250	100	100	100	100	3.51
		0.75	.2512	.185	100	100	100	100	2.71
		0.50	.1709	.121	100	100	100	100	1.81
		0.25	.0873	.059	61	64	80	84	0.95
.75	0.80	1.00	.1660	.156	100	100	100	100	1.74
		0.75	.1323	.123	100	100	100	100	1.34
		0.50	.0947	.088	78	78	97	97	0.94
		0.25	.0519	.048	30	31	54	58	0.53
	1.20	1.00	.1709	.163	100	100	100	100	2.01
		0.75	.1412	.134	100	100	100	100	1.65
		0.50	.1023	.096	89	91	98	100	1.11
		0.25	.0539	.050	34	36	57	62	0.63

1. Null "b" = level of item difficulty for RG.
2. "a" = item discrimination.
3. Diff. "b" = level of difference in "b" between RG and FG.
4. Eff. Size = effect size as in Cohen's typology.
5. Diff. p = difference in corrected p values between RG and FG.
6,7. Signif. level .05, .01 = significance level .05, .01
8,9. MH, LRU = % dif detection rate by procedure
10. Mean Δ_{MH} = mean value for Δ_{MH} under each condition (values are actually negative).

Table 12. MH and LRU DIF detection rates, and Δ_{MH} value, sample size 600.

Null "b"[1]	"a"[2]	Diff "b"[3]	Eff. Size[4]	Diff p[5]	Signif. level .01[6]		Signif. level .05[7]		Δ_{MH}
					MH[8]	LRU[9]	MH[8]	LRU[9]	
-.75	0.80	1.00	.2659	.206	100	100	100	100	2.49
		0.75	.2028	.153	100	100	100	100	1.91
		0.50	.1377	.101	92	93	99	99	1.30
		0.25	.0676	.048	30	34	54	56	0.71
	1.20	1.00	.3291	.250	100	100	100	100	3.47
		0.75	.2512	.185	100	100	100	100	2.70
		0.50	.1709	.121	100	100	100	100	1.79
		0.25	.0873	.059	46	55	74	75	0.97
.75	0.80	1.00	.1660	.156	100	100	100	100	1.73
		0.75	.1323	.123	98	99	100	100	1.36
		0.50	.0947	.088	59	66	84	89	0.94
		0.25	.0519	.048	11	15	30	30	0.45
	1.20	1.00	.1709	.163	100	100	100	100	1.96
		0.75	.1412	.134	100	100	100	100	1.70
		0.50	.1023	.096	86	87	96	98	1.20
		0.25	.0539	.050	28	30	48	50	0.65

1. Null "b" = level of item difficulty for RG.
2. "a" = item discrimination.
3. Diff. "b" = level of difference in "b" between RG and FG.
4. Eff. Size = effect size as in Cohen's typology.
5. Diff. p = difference in corrected p values between RG and FG.
6,7. Signif. level .05, .01 = significance level .05, .01
8,9. MH, LRU = % dif detection rate by procedure
10. Mean Δ_{MH} = mean value for Δ_{MH} under each condition (all values are actually negative).

Table 13. MH and LRU DIF detection rates, and Δ_{MH} value, sample size 400.

Null "b"[1]	"a"[2]	Diff "b"[3]	Eff Size[4]	Diff p[5]	Signif. level .01[6]		Signif. level .05[7]		Δ_{MH}
					MH[8]	LRU[9]	MH[8]	LRU[9]	
-.75	0.80	1.00	.2659	.206	100	100	100	100	2.55
		0.75	.2028	.153	100	100	100	100	2.01
		0.50	.1377	.101	68	79	88	91	1.28
		0.25	.0676	.048	14	17	34	39	0.69
	1.20	1.00	.3291	.250	100	100	100	100	3.55
		0.75	.2512	.185	100	100	100	100	2.71
		0.50	.1709	.121	87	94	96	100	1.78
		0.25	.0873	.059	25	27	41	50	0.92
.75	0.80	1.00	.1660	.156	93	96	98	98	1.62
		0.75	.1323	.123	72	83	88	90	1.34
		0.50	.0947	.088	43	47	60	69	0.93
		0.25	.0519	.048	8	12	29	27	0.50
	1.20	1.00	.1709	.163	98	98	100	100	1.97
		0.75	.1412	.134	95	97	99	95	1.60
		0.50	.1023	.096	57	65	80	89	1.16
		0.25	.0539	.050	17	25	38	44	0.7

1. Null "b" = level of item difficulty for RG.
2. "a" = item discrimination.
3. Diff. "b" = level of difference in "b" between RG and FG.
4. Eff. Size = effect size as in Cohen's typology.
5. Diff. p = difference in corrected p values between RG and FG.
6,7. Signif. level .05, .01 = significance level .05, .01
8,9. MH, LRU = % dif detection rate by procedure
10. Mean Δ_{MH} = mean value for Δ_{MH} under each condition (all values are actually negative.

Table 14. MH and LRU DIF detection rates, and Δ_{MH} value, sample size 200.

Null "b"[1]	"a"[2]	Diff "b"[3]	Eff Size[4]	Diff p[5]	Signif. level .01[6]		Signif. level .05[7]		Δ_{10MH}
					MH[8]	LRU[9]	MH[8]	LRU[9]	
-.75	0.80	1.00	.2659	.206	94	98	98	100	2.52
		0.75	.2028	.153	62	72	84	88	1.87
		0.50	.1377	.101	31	43	59	67	1.33
		0.25	.0676	.048	6	8	15	21	0.67
	1.20	1.00	.3291	.250	100	100	100	100	3.61
		0.75	.2512	.185	92	95	98	99	2.78
		0.50	.1709	.121	53	64	82	89	1.90
		0.25	.0873	.059	8	14	20	30	0.89
.75	0.80	1.00	.1660	.156	55	72	84	89	1.68
		0.75	.1323	.123	30	46	60	70	1.28
		0.50	.0947	.088	16	22	35	44	0.94
		0.25	.0519	.048	6	8	10	18	0.48
	1.20	1.00	.1709	.163	70	84	89	95	1.96
		0.75	.1412	.134	55	62	72	82	1.65
		0.50	.1023	.096	23	33	49	53	1.19
		0.25	.0539	.050	3	6	14	18	0.61

1. Null "b" = level of item difficulty for RG.
2. "a" = item discrimination.
3. Diff. "b" = level of difference in "b" between RG and FG.
4. Eff. Size = effect size as in Cohen's typology.
5. Diff. p = difference in corrected p values between RG and FG.
6,7. Signif. level .05, .01 = significance level .05, .01
8,9. MH, LRU = % dif detection rate by procedure
10. Mean Δ_{MH} = mean value for Δ_{MH} under each condition (values are actually negative).

Table 15. MH and LRU DIF detection rates, and Δ_{MH} value, sample size 100.

Null "b"[1]	"a"[2]	Diff. "b"[3]	Eff Size[4]	Diff. p[5]	Signif. level .01[6]		Signif. level .05[7]		Δ_{MH}
					MH[8]	LRU[9]	MH[8]	LRU[9]	
-.75	0.80	1.00	.2659	.206	57	81	84	100	2.52
		0.75	.2028	.153	18	37	48	88	1.87
		0.50	.1377	.101	10	16	23	67	1.33
		0.25	.0676	.048	1	5	9	21	0.67
	1.20	1.00	.3291	.250	87	96	97	100	3.61
		0.75	.2512	.185	55	70	76	99	2.78
		0.50	.1709	.121	24	34	40	89	1.90
		0.25	.0873	.059	2	3	12	30	0.89
.75	0.80	1.00	.1660	.156	22	37	40	89	1.68
		0.75	.1323	.123	8	21	27	70	1.28
		0.50	.0947	.088	6	13	18	44	0.94
		0.25	.0519	.048	1	2	8	18	0.48
	1.20	1.00	.1709	.163	30	40	48	95	1.96
		0.75	.1412	.134	9	24	34	82	1.65
		0.50	.1023	.096	7	12	18	53	1.19
		0.25	.0539	.050	1	2	9	18	0.61

1. Null "b" = level of item difficulty for RG.
2. "a" = item discrimination.
3. Diff. "b" = level of difference in "b" between RG and FG.
4. Eff. Size = effect size as in Cohen's typology.
5. Diff. p = difference in corrected p values between RG and FG.
6,7. Signif. level .05, .01 = significance level .05, .01
8,9. MH, LRU = % dif detection rate by procedure
10. Mean Δ_{MH} = mean value for Δ_{MH} under each condition (values are actually negative).