INFORMATION TO USERS

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

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

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

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

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

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

UMI®

Université d'Ottawa · University of Ottawa

DETECTING DIF IN POLYTOMOUS ITEM RESPONSES

By

Fang Tian

Dissertation presented to the School of Graduate
Studies and Research as partial fulfillment
of the Ph.D. degree in Education

Faculty of Education, University of Ottawa
Ottawa, Canada, 1999

Acknowledgments

I would like to acknowledge with appreciation the contributions of all those who helped to make the completion of this dissertation possible.

I would especially like to thank my advisor, Dr. Marvin W. Boss for all his time, support, expertise, and suggestions through the process of this dissertation. I would like to take this opportunity to thank him for not only being there as an advisor but also as a friend.

I would also like to thank Dr. Marc Gessaroli for his support and constructive comments over the phone and via email over the past four years.

I would also like to thank Dr. Judith Spray for providing software and relevant documentation.

Finally, a special thank you goes to Sharon and Ying. I would not have completed this thesis without their patience, understanding, and support.

TABLE OF CONTENTS

LIST OF TABLES

Abstract

Use of polytomously scored items in educational tests is becoming increasingly common with new types of assessments such as performance assessment. It is believed that performance assessment provides more equitable approaches to testing than traditional multiple-choice tests. However, some forms of performance assessment may be more likely than conventional tests to be influenced by construct-irrelevant factors, resulting in differential item functioning (DIF). Several methods have been proposed for DIF assessment with polytomous item responses, such as the Mantel procedure (Mantel, 1963), the generalized Mantel-Haenszel procedure (GMH) (Mantel & Haenszel, 1959), the polytomous extension of the standardization approach (STND) (Potenza & Dorans, 1995), the polytomous extension of SIBTEST (SIBTEST) (Chang, Mazzeo, & Roussos, 1996), and logistic discriminant function analysis (LDFA) (Miller & Spray, 1993). The results of previous studies on the performance of the above procedures are very encouraging. However, the conditions under which different procedures were examined are limited. The utilities of these procedures in assessing DIF in polytomous items have not received the thorough and rigorous study accorded to their dichotomous DIF counterparts. Further research is needed to investigate their performance.

The purpose of this study was to investigate and compare the performance of the Mantel, GMH, STND, SIBTEST, and LDFA (LDFA-Uniform and LDFA-Nonuniform) procedures in detecting DIF with polytomous item responses under a variety of conditions. A simulation study was conducted to evaluate their Type I error and power in detecting DIF when the properties of the items were known. The factors considered were

group ability distribution differences (three levels), test length (two levels), sample size (three levels), sample size ratio (two levels), item discrimination (null DIF, constant uniform DIF, and balanced uniform DIF) or item-discrimination parameter difference between reference and focal groups (nonuniform DIF) (three levels), and DIF conditions (four levels). Data sets in each of the 3x2x3x2x3x4 conditions were replicated 100 times. This yielded a total of 43,200 data sets. Each data set was, in turn, analyzed with the each DIF detection procedure. The effects of the manipulated factors on Type I error rates and power of DIF detection procedures were analyzed using logit analysis. The outcome (i.e. reject or do not reject at α=.05) was used as the dependent variable. The independent variables were test length, sample size, sample size ratio between the focal and reference groups, and studied-item *a*-parameter or *a*-parameter difference. The results were analyzed for each procedure and for each level of group ability difference separately.

Based on the findings of this study, all procedures had good and comparable Type I error control when groups were equal in ability or when groups had a small ability difference and item discrimination was not high or when groups had a large ability difference and item discrimination was low. The Type I error of all procedures was generally unacceptable with the combination of small group ability difference and high discrimination and the combination of large group ability difference and medium to high discrimination. The power of these procedures under these conditions was questionable.

All procedures (except LDFA-Nonuniform) investigated in this study were very powerful in detecting constant uniform DIF, with the LDFA-Uniform being the most powerful, followed by the Mantel, SIBTEST, STND, and GMH.

The Mantel, STND, SIBTEST, and LDFA-Uniform were not useful for detecting balanced uniform DIF and nonuniform DIF. The GMH and LDFA-Nonuniform can be used to detect balanced uniform DIF with GMH being much more powerful. However, GMH was not very powerful when sample size was low or when item discrimination was low; the power of LDFA-Nonuniform was adequate only when item discrimination was high and sample size was large.

Both LDFA-Nonuniform and GMH can be used to detect nonuniform DIF with LDFA-Nonuniform being more powerful when two groups had equal ability distributions and GMH being more powerful when two groups had different ability distributions. However, LDFA-Nonuniform was not very powerful when sample size was low or when item discrimination was low. The power GMH was adequate only when item discrimination was high and sample size was large.

Regarding sample size, it appeared that sample sizes of 600/600 and 400/800 produced a balance between reasonable Type I error and adequate power. With respect to sample size ratio, both ratios used in this study worked well. However, for STND, the ratio of 1:2 produced relatively lower Type I error and consequently lower power than the ratio of 1:1. As for test length, both test lengths used in this study seemed to be appropriate. However, the longer test produced slightly lower Type I error rates and higher power. In terms of studied-item discrimination, except for SIBTEST, the Type I error of all the other DIF detection procedures in this study seemed to be within the theoretical nominal value when item discrimination was low but was seriously inflated at high item discrimination when there was a group ability difference.

Finally, results from this study indicated that the LDFA procedure is preferable over the other procedures. It had good adherence to Type I error when two groups had equal ability distributions. Most important of all, it can distinguish between uniform and nonuniform DIF and can be used to detect constant uniform DIF, balanced uniform DIF, and nonuniform DIF. However, GMH may be preferable for balanced uniform DIF.

CHAPTER I
INTRODUCTION

Statement of the Problem

The validity of a test may be affected by many factors. Among those factors, differential item functioning (DIF) is a serious one that threatens test validity. DIF indicates a difference in item performance between two comparable groups of examinees that are matched with respect to the construct being measured by the test (Potenza & Dorans, 1995). Different groups of examinees may react differently to the same test questions. By investigating these differences, one may better understand both the characteristics of test items and the characteristics of different groups of examinees in terms of their cognitive processes, test-taking strategies, and knowledge deficiencies. Test developers and users have long recognized that test scores may be influenced by extraneous factors other than the construct a test was designed to measure. Because of these factors, observed differences in test scores do not always accurately reflect differences in examinees' true abilities on the construct measured by the test. Sometimes, certain examinees are prevented from showing their true abilities by the way test questions are presented. For example, negatively-worded questions may contain a source of difficulty not relevant to the test construct. Tests containing such items may have reduced validity for comparing different groups of examinees, because inter-group score difference may be indicative of an irrelevant construct. Bias in individual test questions may be thought of as systematic error that distorts the meaning of test inferences for members of a particular group (Camilli & Shepard, 1994). It is important to ensure that

the items in a test do not convey undue advantage for success to the members of one subgroup over those of another.

Traditionally, multiple-choice questions have been the most widely used item format in educational achievement tests. For the vast majority of multiple-choice tests, items are scored dichotomously (i.e., correct or incorrect). Even when a multiple-choice test is not scored this way, it is often analysed as if it were (Potenza & Dorans, 1995). Educational reform efforts have led to an increased use of alternatives to the traditional dichotomously scored multiple-choice items (Potenza & Dorans, 1995). It is believed that new types of assessment forms such as performance assessment provides more equitable approaches to testing which yield more information than multiple-choice measures. However, some forms of performance assessment may in fact be more likely than conventional tests to be influenced by construct-irrelevant factors (Zwick, Donoghue, & Grima, 1993), resulting in DIF. For these new item types, the outcome of interest is not simply a 1 or 0 but a judgment of the quality of a writing portfolio, a mathematical proof, or a scientific experiment. Item responses to these new item types usually consist of more than two categories which may be scored on an ordinal scale and are usually termed *polytomous item responses*. Assessing DIF in polytomous items with ordered response categories is the primary interest in the current study

Many current psychometric procedures were developed primarily for analysis with dichotomously scored, multiple-choice test items, and they do not easily accommodate polytomously scored item responses. The demand for the analysis of polytomous item responses is increasing due to the increased popularity of performance

assessment. Several item response theory (IRT) models, such as the graded response model (Samejima, 1969), the partial credit model (Masters, 1982), the rating scale model (Andrich, 1978), the nominal response model (Bock, 1972), and the generalized partial credit model (Muraki, 1992) have been developed to model polytomous data with different scoring formats.

There are numerous methods for conducting DIF assessment for dichotomously scored items. However, there is not yet a consensus about how to conceptualize, measure, or test DIF when item responses are polytomously scored. With dichotomous item responses, the most widely used DIF detection methods are procedures based on item response theory (Thissen, Steinberg, & Wainer, 1988), SIBTEST (Shealy & Stout, 1993), logistic regression (Swaminathan & Rogers, 1990), standardization (Dorans & Kulick, 1986), and the Mantel-Haenszel procedure (Holland & Thayer, 1988). These methods have been studied extensively and rigorously. They have gained wide acceptance as useful DIF detection methods for dichotomous item responses. Several extensions of the dichotomous DIF procedures have been proposed for use with polytomous item responses, such as the polytomous logistic regression procedure (French & Miller, 1996), the Mantel procedure for ordered response categories (Mantel, 1963), the generalized Mantel-Haenszel procedure for nominal data (Mantel & Haenszel, 1959), the polytomous extension of the standardization approach (Potenza & Dorans, 1995), the polytomous extension of SIBTEST (Chang, Mazzeo, & Roussos, 1996), and logistic discriminant function analysis (Miller & Spray, 1993). However, their utilities in assessing DIF in polytomous items have not received the thorough and rigorous study accorded to their

dichotomous DIF counterparts. Further research is needed to investigate their performance before they are ready for routine operational use.

Purposes of the Study

The purpose of this study was to assess and compare the performance of several polytomous DIF detection procedures under various conditions of group ability distribution difference, test length, sample (sample size and sample size ratio between focal and reference groups), studied-item discrimination parameter, and type of DIF. The procedures of interest in this study are: the Mantel procedure (Mantel, 1963), the generalized Mantel-Haenszel procedure (GMH) (Mantel & Haenszel, 1959), the polytomous extension of the standardization approach (STND) (Potenza & Dorans, 1995), the polytomous extension of SIBTEST (SIBTEST) (Chang, Mazzeo, & Roussos, 1996), and logistic discriminant function analysis (LDFA) (Miller & Spray, 1993).

In the next chapter, research studies relevant to the purpose of this study are reviewed and specific research questions for this study are formulated.

Organization of This Work

In Chapter 2, DIF is defined and statistical modelling of DIF from different perspectives is presented. The terminology associated with DIF study is presented. This is followed by a review of the techniques that have been proposed for assessing DIF with polytomous item responses and the empirical performance of these procedures. Then, relationships among the polytomous DIF detection procedures are explored, and the need

to assess the performance of polytomous DIF procedures is discussed. Finally, research questions are formulated.

In Chapter 3, the methodology used to assess the performance of DIF detection techniques is presented. Included are the conditions under which the statistics were investigated, item parameters used, data generation procedures, software used, and data analysis methods.

In Chapter 4, the results of this investigation are presented for each procedure.

In Chapter 5, the results of this study are discussed. The procedures are compared in their effectiveness in assessing DIF.

Finally, in Chapter 6, the conclusions which can be drawn from this study are presented, limitations associated with this study are discussed. Suggestions for future research are also provided in this chapter.

CHAPTER II
REVIEW OF LITERATURE

In this chapter, a review of relevant literature is presented. First, a definition of differential item functioning (DIF) is given and its historical origin is explored. This is followed by a presentation of statistical modelling of DIF for both dichotomous and polytomous item responses. Next, the polytomous extensions of some dichotomous DIF detection procedures are described and studies on the utilities of these procedures in detecting DIF in polytomous items are reported. Then, a summary of the studies of polytomous DIF procedures is provided and the need to assess the performance of polytomous DIF procedures as was done in this study is discussed.

DIF: Definition and History

The study of test items that function differently for different groups of examinees was originally called "item bias" research (Holland & Thayer, 1986, 1988). The process of investigating item bias involves gathering evidence of the relative performance of different groups of examinees on a test item. Evidence of differential performance is a necessary, but not sufficient basis, for the conclusion that a test item is biased. This conclusion rests additionally on an inference that goes beyond the data (Hambleton, Swaminathan, & Rogers, 1991). The empirical indices of bias do not necessarily indicate bias in the sense of unfairness. Groups may differ in their responses to an item for such other reasons as differential course taking patterns or interest in the specific content of the item. Whenever items appear to be measuring differently for different subgroups of

examinees, test developers must conduct further investigation to determine whether the source of differential difficulty is relevant or irrelevant to the construct being measured by the test. Only in the latter case can the items be said to be biased against members of the affected group. From empirical evidence, a valid conclusion that could be made about an item is that it is functioning differentially with different groups of examinees. A more neutral term to "item bias", differential item functioning (DIF), is preferred to describe the empirical evidence obtained in the investigation of bias (Hambleton, Swaminathan, & Rogers, 1991). DIF investigation procedures are often used as a part of test development to detect possibly flawed items. To support conclusions about possible bias in test items, further effort then has to be made to obtain substantive explanations for DIF.

The psychometrically accepted definition of DIF is that an item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item correct (Hambleton, Swaminathan, & Rogers, 1991). This requires that the individuals in the groups being compared are of comparable ability, as measured by a test or other means. If the item performance of two unmatched groups of examinees is being compared, the result may be a measure of item impact rather than of DIF (Holland & Thayer, 1986, 1988).

DIF can be further distinguished as uniform or nonuniform. An item displays uniform DIF if there is no interaction between ability level and group membership and the probability of correctly answering the item is greater for one group than the other uniformly over the entire ability range. Nonuniform DIF exists when there is an interaction between ability level and group membership; the difference in the probability

of answering the item correctly for the two groups is not the same across all ability levels (Swaminathan & Rogers, 1990). The advantage may even shift from one group to the other, depending on ability level. In terms of item response theory (IRT), uniform DIF is represented by parallel but not coincident item characteristic curves (ICCs) while nonparallel or crossing ICCs indicate nonuniform DIF. An ICC is a monotonically increasing function that describes the relationship between examinees' item performance and the set of traits underlying item performance. This function specifies that as the level of the trait increases, the probability of a correct response to an item increases (Swaminathan & Rogers, 1990).

There seems to be a consensus among the testing community that DIF is caused by the multidimensionality of the test together with differences between groups in the underlying multidimensional ability distributions. DIF may even be thought of as a special type of multidimensionality. However, multidimensionality in itself is not a sufficient condition for the occurrence of DIF. DIF occurs only when an item measures one or more dimensions in addition to the abilities common to all items and only when subgroups differ in their ability distributions on the secondary dimension(s) (Camilli & Shepard, 1994). An item is considered free of DIF if examinees at the same level in the common ability have an equal probability of answering the item correctly, regardless of group membership. This definition implies that any observed DIF is due to multidimensionality. If the item and the total test score were measuring the same common ability, or exactly the same weighted composite of common abilities, then it would be impossible for differences in item performance to exist for examinees with the

same ability or abilities except due to chance. If there are significant group differences in item performance, then it must be the case that something other than the construct to be measured by the test is influencing performance on that item. Therefore, the test must be multidimensional, and it is this multidimensionality together with differences between groups in the underlying multidimensional ability distributions, which results in DIF (Mazor, Hambleton, & Clauser, 1994).

Statistical DIF Modelling

In the following, DIF modeling for dichotomous items is first reviewed, then a definition of DIF for polytomously scored items is presented.

Dichotomous DIF Modelling

Normally, in a DIF study, the item performances of two groups are compared. The group suspected of being disadvantaged, usually the minority group, is called the focal group, and the group against which the performance of the focal group is compared, usually the majority group is called the reference group. The item under investigation is called the studied item. DIF can be modelled using either an observed score, or an estimate of a latent variable, or an estimate of true score as the conditional variable (Chang et al., 1996).

Latent-Variable Null DIF

Let Y be the score on the studied item and θ be the latent matching variable. Denote $E_R[Y|\theta]$ and $E_F[Y|\theta]$ as regressions of Y on θ for reference and focal groups, respectively. An item does not exhibit DIF if, for all values of θ,

$$E_R[Y|\theta] = E_F[Y|\theta]. \quad (1)$$

An item exhibits DIF if the regressions of Y on θ are not identical for the groups under study (Chang et al., 1996).

Observed-Score Null DIF

Let Y be the score on the studied item and X be the observed matching test score, including the score on the studied item. Denote $E_R[Y|X]$ and $E_F[Y|X]$ as regressions of Y on X for reference and focal groups, respectively. An item does not exhibit DIF if, for all values of X,

$$E_R[Y|X] = E_F[Y|X]. \quad (2)$$

An item exhibits DIF if the expected scores are not the same for the reference and focal groups matched on X (Chang et al., 1996).

Classical Test Theory Latent-Variable Null DIF

Let $E_R[Y|t]$ and $E_F[Y|t]$ denote the regressions of Y on the matching test true score t for reference and focal groups, respectively. An item does not exhibit DIF if, for all values of t,

$$E_R[Y|t] = E_F[Y|t]. \quad (3)$$

In other words, an item exhibits DIF if $E_R[Y|t] \neq E_F[Y|t]$ (Chang et al., 1996).

Polytomous DIF Modeling

Chang et al. (1996) extended the dichotomous DIF modeling to the polytomous item responses. Let Y be the score on the studied item. Assume Y can be scored in terms of m+1 ordered categories (e.g., $Y = k$, $0 \leq k \leq m$). Let $P_{k.G}(\theta)$ denote the item category response function (ICRF), the probability of getting score k for a randomly sampled

examinee with ability θ from group G (G = R for the reference group or F for the focal group). The regression of item score on latent ability can be defined as a weighted sum of item category response functions ICRFs:

$$E_G[Y|\theta] = \sum_{k=1}^{m} kP_{k,G}(\theta). \tag{4}$$

As with the dichotomous items, this regression is referred to as an *item response function* (IRF) in which dichotomous items can be viewed as a special case with m = 1 (Chang et al., 1996).

A latent-variable definition of null DIF for polytomous items can be directly generalized from Equation 1. That is, an item does not exhibit DIF if the regression of polytomous item score on the latent variable (which is now the weighted sum of ICRFs) is identical for the reference and focal groups, i.e., $E_R[Y|\theta] = E_F[Y|\theta]$. Likewise, an observed-score definition of null DIF and a classical test theory true-score definition of DIF can be directly generalized from their dichotomous counterparts. It is obvious that the generation of the definition of polytomous null DIF from dichotomous modeling of null DIF can be justified from the item structure invariance perspective.

Unlike dichotomous IRT models in which the structure of an item is completely determined by the specification of its IRF, for polytomous IRT models, the item structure is completely determined only if all m ICRFs are specified. A unique correspondence may or may not exist between an IRF and a set of ICRFs. If Equation 1 does not in general imply

$$P_{kR}(\theta) = P_{kF}(\theta), \qquad k = 1,..,m, \tag{5}$$

for all θ, where $P_{kR}(\theta)$ and $P_{kF}(\theta)$ are ICRFs for the reference and focal groups, respectively, then Equation 5 might be used as the latent-variable null-DIF definition. Chang and Mazzeo (1994) have proven that Equation 1 does indeed imply Equation 5 for three of the most commonly used polytomous item response models—partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1992), and the graded item response model (Samejima, 1969). DIF in polytomous items can be assessed only with respect to differences in IRFs across ability groups, without considering differences in ICRFs. No information is lost by comparing only the IRFs (Chang et al., 1996).

Descriptions of Polytomous DIF Detection Procedures Studied

Use of polytomously scored items in educational tests is becoming increasingly common with new types of assessments such as performance assessment and portfolio analysis. This has created the need for statistical DIF procedures that can accommodate polytomously scored item responses. Within the IRT framework, several models can be used to study polytomous DIF, such as the graded response model (Samejima, 1969), partial credit model (Masters, 1982), and generalized partial credit model (Muraki, 1992). The model comparison approach uses a series of likelihood ratio tests to assess the significance of DIF effects, contrasting a compact model, in which focal and reference group item category response functions are identical, with different augmented models in which the item category response functions differ. As with the dichotomous IRT-based procedures, difficult problems sometimes arise in applying the polytomous IRT-based

DIF procedures. Strict IRT assumptions, complex statistical calculation, large sample size requirement, and poor parameter estimation often make the implementation of IRT-based methods problematic, if not impossible, in some situations. It seems more reasonable and practical to resort to non-IRT methods. Several non-IRT procedures have been proposed to assess DIF in polytomous items; some are extensions of dichotomous DIF procedures. These procedures are the polytomous logistic regression procedure (French & Miller, 1996), the Mantel procedure for ordered response categories (Mantel, 1963), the generalized Mantel-Haenszel procedure for nominal data (GMH) (Mantel & Haenszel, 1959), the polytomous extension of the standardization approach (STND) (Potenza & Dorans, 1995), the polytomous extension of SIBTEST (SIBTEST) (Chang, Mazzeo, & Roussos, 1996), and logistic discriminant function analysis (LDFA) (Miller & Spray, 1993). In the following, these procedures are described in detail.

Extensions of the Mantel-Haenszel Procedure to Polytomous Data

The Mantel-Haenszel (MH) procedure (Mantel & Haenszel, 1959) is one of the most popular DIF detection methods for dichotomous items. When applying the MH procedure to study DIF in dichotomous items, groups are first matched on a measure of ability, usually the total test score. Inclusion of the studied item in the matching variable in the dichotomous case has been supported by several researches (Donoghue, Holland, & Thayer, 1993; Holland & Thayer, 1986, 1988; Zwick, 1990). The data for the studied item for the examinees in the reference and focal groups are arranged into a series of 2x 2 contingency tables, one for each level of the matching variable. The null DIF hypothesis is that the odds for getting the item correct are the same for the reference and focal

groups, given a level of the matching variable, across all levels of the matching variable. There is a one-degree-of-freedom chi-square test associated with the MH approach. Let α denote the ratio of the odds of answering the item correctly for the reference group to the corresponding odds for the focal group; the $MH\chi^2$ tests the null hypothesis of H_0: $\alpha = 1$ against H_1: $\alpha \neq 1$. The estimate of the MH common odds ratio is given by α_{MH}, which ranges from 0 to ∞ with a value of 1 indicating null DIF. Using a log-odds transformation, Holland and Thayer (1986) converted α_{MH} into a difference on a delta (Δ) scale, called $MH_{D\text{-}DIF}$. $MH_{D\text{-}DIF}$ has been frequently used as a measure of DIF.

Two generalizations of the dichotomous MH procedure can be applied to assess DIF in polytomous item responses (Zwick, et al., 1993), one for ordinal response data, the other for nominal response data.

The Ordinal Version of the Extended MH Statistic (Mantel)

Mantel (1963) proposed an extension of the MH procedure for ordered response categories that involves comparing the means between matched groups. To apply the Mantel procedure in the DIF context, the reference and focal groups should first be matched on a measure of ability, as with the dichotomous MH procedure. The data are arranged into a series of $2 \times K$ contingency tables as shown in Table 1, one for each score level of the matching variable.

Table 1

Data for the *m*th Level of the Matching Variable

	Score on Studied Item					
Group	Y_1	Y_2	Y_3		Y_K	Total
Reference	N_{Rm1}	N_{Rm2}	N_{Rm3}	...	N_{Rmk}	N_{Rm}
Focal	N_{Fm1}	N_{Fm2}	N_{Fm3}	...	N_{Fmk}	N_{Fm}
Total	N_{m1}	N_{m2}	N_{m3}	...	N_{mk}	N_m

where

M is the number of the levels of the matching variable, $m = 1, \ldots, M$,

K is the number of response categories, $k = 1, \ldots, K$,

$Y_1, Y_2, \ldots, Y_K$, represent the K scores that can be obtained on the studied item,

N_{Rmk} and N_{Fmk} denote the number of reference and focal group members, respectively, who are at the *m*th level of the matching variable and received an item score of Y_k,

N_{Fm} and N_{Rm} denote the total number of focal and reference group members, respectively, at the *m*th level,

N_m is the total number of examinees at the *m*th level.

Mantel (1963) proposed a one degree-of-freedom test of no conditional association (H_o) between the item score and group membership for the studied item given a fixed value of the matching variable:

$$Mantel\chi^2 = \frac{\left(\sum_m F_m - \sum_m E(F_m)\right)^2}{\sum_m Var(F_m)}, \tag{6}$$

where

$$F_m = \sum_k Y_k N_{Fmk}, \tag{7}$$

$$E(F_m) = \frac{N_{Fm}}{N_m} \sum_k Y_k N_{mk}, \tag{8}$$

$$Var(F_m) = \frac{N_{Rm} N_{Fm}}{N_m^2 (N_m - 1)} \left\{ \left(N_m \sum_k Y_k^2 N_{mk} \right) - \left(\sum_k Y_k N_{mk} \right)^2 \right\}. \tag{9}$$

Under H_0, $Mantel\chi^2$ is distributed as chi-square with one degree of freedom. Rejecting H_0 indicates that members of the reference and focal groups who have the same ability differ in their mean performance on the studied item.

The Nominal Version of the Extended MH Statistic (GMH).

Mantel and Haenszel (1959) introduced a generalized MH statistic (GMH) for nominal response data that involves comparing the entire response distributions between matched groups. The test statistic for the GMH procedure is a multivariate generalization of the MH chi-square statistic for the dichotomous case (Zwick, et al., 1993). In the notation of Table 1,

$$A'_m = (N_{Rm1}, N_{Rm2}, \ldots N_{Rm(k-1)}), \tag{10}$$

$$E(A'_m) = N_{Rm} N'_T / N_m, \tag{11}$$

$$N'_T = (N_{m1}, N_{m2}, \ldots, N_{m(k-1)}), \tag{12}$$

$$V(A_m) = N_{Rm} N_{Fm} \left(\frac{N_m dia(N_T) - N_T N_T'}{N_m^2 (N_m - 1)}\right), \tag{13}$$

and $dia(N_T)$ is a (k-1) by (k-1) diagonal matrix with elements N_T. Whereas A_m, $E(A_m)$, and $V(A_m)$ are scalars in the dichotomous case, A_m and $E(A_m)$ are now vectors of length k-1, corresponding to k-1 of the k response categories, and $V(A_m)$ is a (k-1) by (k-1) covariance matrix. Then the GMH test statistic is given by

$$GMH\chi^2 = \left[\sum A_m - \sum E(A_m)\right]' \left[\sum V(A_m)\right]^{-1} \left[\sum A_m - \sum E(A_m)\right] \tag{14}$$

This statistic has a large sample chi-square distribution with k-1 degrees of freedom under the null hypothesis of conditional independence between group membership and item response. A significant test statistic implies that uniform DIF is present in the item. The GMH statistic does not explicitly take into account the possible ordering of response categories; instead, it provides for the comparison of the two groups in terms of their entire response distributions, rather than their means alone. The odds that focal group members will be assigned a particular score category can be compared to the odds for the reference group, conditional on the matching variable (Zwick, el al., 1993).

Polytomous Extension of the Standardization Approach (STND)

Closely related to the dichotomous MH procedure is the Standardization (SNTD) approach (Dorans & Kulick, 1986). STND defines an item as not showing DIF when the expected performance on an item is the same for examinees of equal ability from different groups. This definition has been shown to be equivalent to the definition of null DIF for the MH procedure (Dorans & Holland, 1993; Potenza & Dorans, 1995). Expected performance on an item can be operationalized by nonparametric item test regressions.

Differences in empirical item test regressions are indicative of DIF. An average overall index of DIF, $STND_{P\text{-}DIF}$ (Dorans & Holland, 1993), is obtained by averaging differences in expected item scores (proportion correct) across levels of the matching variable, weighting each difference by focal group relative frequencies. The STND approach is a very useful way to describe DIF and has been shown to perform well in practical applications (Donoghue, et al., 1993).

Analogous to the $STND_{P\text{-}DIF}$ for the case of dichotomous items, a general index, $STND_{ES\text{-}DIF}$, or standardized expected item score DIF, was developed for polytomous data (Potenza & Dorans, 1995). The only difference between the two is the number of levels of the dependent variable (i.e., the item score). The $STND_{P\text{-}DIF}$ compares the means of the reference and focal groups, adjusting for differences in the distribution of reference and focal group members across the levels of the matching variable. Let X be the matching variable with M levels, $m = 1, \ldots, M$, and let Y be the ordered item score with K categories, $k = 1, \ldots, K$. For the polytomous version of STND, first compute expected item scores for both the reference and focal groups—$E_{Fm}(Y|X)$ and $E_{Rm}(Y|X)$—using

$$E_{Fm}(Y|X) = \sum_k N_{Fmk} Y_k \; / \; N_{Fm} \tag{15}$$

and

$$E_{Rm}(Y|X) = \sum_k N_{Rmk} Y_k \; / \; N_{Rm}, \tag{16}$$

where N_{Fmk}, N_{Rmk}, N_{Fm}, N_{Rm} have already been defined in Table 1. The item score variable, Y_k, can take on any ordered value, including 1, 2, 3, ..., K.

Then the differences in expected item score at each level of the matching variable can be computed,

$$D_m = E_{Fm}(Y|X) - E_{Rm}(Y|X). \tag{17}$$

A general index can be obtained by weighting these differences by focal group relative frequencies (Dorans & Kulick, 1986):

$$STND_{ES-DIF} = \sum_m N_{Fm} D_m / N_F, \tag{18}$$

where N_F is the total number of focal group examinees. A negative $STND_{ES\text{-}DIF}$ value implies that, conditional on the matching variable, the focal group has a lower mean item score than the reference group.

Both the polytomous extension of the STND statistic and the Mantel (Mantel, 1963) procedure emphasize expected (average) item scores when comparing reference and focal groups. Also for both procedures, the matching variable must include the studied item in order to produce an unbiased estimate (Donoghue, et al., 1993; Holland & Thayer, 1988). Since there is no descriptive statistic analogous to α_{MH} for the Mantel procedure (Mantel, 1963). $STND_{ES\text{-}DIF}$ can be used as a descriptive statistic to supplement the Mantel (Mantel, 1963) approach to make the results of the latter more interpretable. The expression in Equation 18 is defined as the standardized mean difference (SMD) by Zwick et al. (1993).

Zwick and Thayer (1996) developed two standard error formulae for $STND_{ES\text{-}DIF}$ (called SMD in their paper). One is based on Mantel's (1963) hypergeometric model. The other is based on a multinomial model. Studies (Zwick & Thayer, 1996; Zwick, Thayer, & Mazzeo, 1997) indicated that the standard error based on Mantel's (1963)

hypergeometric model performed better. Therefore, it is described in the following and is used in this study. The variance formula for $STND_{ES\text{-}DIF}$, developed by Zwick and Thayer (1996), reformulated in the notation used here, is defined as follows:

$$Var(STND_{ES-DIF}) = \sum_m \left(\frac{N_{Fm}}{N_m}\right)^2 \left(\frac{1}{N_{Fm}} + \frac{1}{N_{Rm}}\right)^2 Var(F_m), \tag{19}$$

where N_{Fm} and N_m have been defined in Table 1 (p.15) and $Var(F_m)$ is defined by Equation 9.

A test statistic $Z(STND_{ES\text{-}DIF})$ is formed by dividing $STND_{ES\text{-}DIF}$ by its standard error:

$$Z(STND_{ES-DIF}) = \frac{STND_{ES-DIF}}{\sqrt{Var(STND_{ES-DIF})}}. \tag{20}$$

The statistic $Z(STND_{ES\text{-}DIF})$ is approximately distributed as standard normal variates under H_o. The standardization approach provides both a descriptive measure of DIF and the basis for a test of the hypothesis of no conditional association between group membership and item score.

Polytomous Extension of SIBTEST Procedure

Shealy and Stout (1993) introduced SIBTEST to assess DIF in dichotomous items. The null DIF definition for SIBTEST is that an item exhibits DIF if the expected scores are identical for the reference and focal groups matched on θ. The amount of DIF at θ is measured by

$$B_0(\theta) \equiv E_R[Y|\theta] - E_F[Y|\theta]. \tag{21}$$

A global index of DIF was defined by Shealy and Stout (1993) for the dichotomous case as

$$\beta = \int B_0(\theta) f_F(\theta) d\theta, \tag{22}$$

where $f_F(\theta)$ denotes the density of θ in the focal group. The global index β can be interpreted as the expected amount of DIF experienced by a randomly selected focal group examinee. SIBTEST then tests H_o: $\beta = 0$ versus H_l: $\beta \neq 0$.

SIBTEST is a nonparametric latent-variable procedure (Potenza & Dorans, 1995). Although SIBTEST adopts a latent-variable definition of DIF, this procedure neither requires nor uses IRT ability or item parameter estimates for its calculation. SIBTEST assesses DIF in the same manner as STND, except that instead of using the empirical item-test regression as used in STND, SIBTEST regresses item performance onto an estimate based on classical test theory of matching variable true score.

SIBTEST was designed to examine unidirectional DIF. Li and Stout (1995) modified the SIBTEST and introduced Crossing SIBTEST that tests for crossing DIF by first estimating the matched subtest score at which the crossing takes place and then estimates the absolute difference between $E_R(Y|\theta)$ and $E_F(Y|\theta)$ across all θ. Unidirectional DIF occurs when the probability of answering an item correctly for one group is greater than the other over the entire ability range, regardless of whether the magnitude of DIF is constant or not constant as ability level varies. Crossing DIF exists when the difference in the probability of getting an item right for the two groups changes sign over the ability range. In IRT terms, crossing DIF is indicated by two crossing ICCs while unidirectional DIF is displayed by two non-crossing ICCs.

From the perspective of item structure invariance, Equations (1), (21), and (22) form an equally appropriate basis for testing DIF in both dichotomous and polytomous items. Based on this, Chang, et al. (1996) made some modifications to the dichotomous SIBTEST procedure. The modified procedure includes dichotomous item scoring as a special case. In using the modified SIBTEST, the studied item can be either dichotomous or polytomous, and the matching test can consist of a mixture of both types of items.

In deriving the formulae for polytomous SIBTEST, the following notations are given by Chang et al. (1996):

Y is the studied item score which has $k + 1$ ordered categories (i.e., $Y = 0, 1, \ldots, k$).

$X_1, X_2, \ldots X_n$ are the item scores for the n matching items.

$Y_1, Y_2, \ldots, Y_n$ are the maximum possible scores for $X_1, X_2, \ldots, X_n$, respectively. For example, $X_j = 0, 1, \ldots, Y_j$; if $Y_j = 1$, then X_j is dichotomously scored.

$X = \sum_{j=1}^{n} X_j$ is the matching score. $X = 0, 1, .., n_H$, where

$$n_H \equiv \sum_{j=1}^{n} Y_j \tag{23}$$

is the maximum possible matching score.

$\overline{Y}_{gm}$ is the average score on the studied item for all group G (G = F or R) examinees for whom $X = m$.

p_{Fm} is the proportion of examinees in the focal group getting score $X = m$ on the matching test $X_1, \ldots, X_n$. That is, $p_{Fm} = N_{Fm}/N_F$, where N_F is the total number of focal

group examinees, and N_{Fm} is the total number of focal group examinees with matching test score $X = m$.

p_m is the proportion of examinees getting score $X = m$ on the matching test $X_1, \ldots, X_n$. That is, $p_m = N_m/N$, where N is the number of total examinees, and N_m is defined by $N_m = N_{Rm} + N_{Fm}$.

In deriving formulae for estimating and testing DIF in polytomous items, Chang et al. (1996) made use of the classical test theory latent-variable definition of DIF. Chang (1994) proved that under the partial credit model or generalized partial credit model, matching test true scores are simply a one-to-one transformation of the latent variable θ. Analogously to Equation 21, the local measure of DIF at the matching test true score t can be defined as $B(t) = E_R[Y|t] - E_F[Y|t]$. Thus, the corresponding DIF index can be defined as $\beta = \int B(t) f_F(t) dt$. In the unidimensional IRT case, β is the same as in Equation 22.

DIF could be estimated locally by

$$d_m = \bar{Y}_{Rm} - \bar{Y}_{Fm}, \qquad m = 0,\ldots,n_H, \tag{24}$$

where $\bar{Y}_{Rm} - \bar{Y}_{Fm}$ is the group difference in performance on the studied item among examinees with the same observed matching test score. Assuming that examinees with the same matching test *observed* score have (approximately) the same *true* score, the d_m is also approximately the difference in item scores at the same true score. If the studied item does not have observed-score DIF, then $d_m \approx 0$. Chang et al. (1996) also suggested a

statistic to estimate the DIF β for the special case where the reference and focal groups have the same θ distribution is

$$\beta = \sum_{m=0}^{n_H} p_m d_m. \tag{25}$$

A test statistic can be defined by

$$B = \frac{\beta}{\sigma(\beta)}, \tag{26}$$

where

$$\sigma(\beta) = \left[\sum_{m=0}^{n_H} p_m^2 \left(\frac{\sigma^2(Y|m,R)}{N_{Rm}} + \frac{\sigma^2(Y|m,F)}{N_{Fm}} \right) \right]^{1/2}, \tag{27}$$

and $\sigma^2(Y|m,G)$ is the sample variance of the studied item scores for examinees in group G (G = R or F) with a total score $X = m$ on the matching test.

To correct the highly inflated impact-induced Type I error rates of the polytomous SIBTEST, Chang et al. (1996) used the regression correction method used by Shealy and Stout (1993) for the dichotomous SIBTEST. However, they modified the matching test reliability estimates used by Shealy and Stout in their regression correction, substituting the more general coefficient α for the dichotomous-item-based *KR*-20.

Extensions of Logistic Regression Procedure to Polytomous Data

The logistic regression (LR) French and Miller (1996)procedure is not of interest in this study for reasons explained later. However, it is closely related to the logistic discriminant function analysis which is described in the next section and which is used in this study. Therefore, a general description is given here. Details of how the LR procedure can be applied to polytomous data can be found in Agresti (1990) and French

and Miller (1996).

The LR procedure for the dichotomous case is an observed score method that specifies a particular parametric form for the relationship between item score and matching variable (Potenza & Dorans, 1995). Studies have shown that the logistic regression procedure is powerful in detecting both uniform and nonuniform DIF in dichotomous items (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). The dichotomous LR procedure models the probability of observing each dichotomous item response, U, as a function of two explanatory variables: observed test score, X, and a group indicator variable, G. The LR model (without item notation) can be written (Hosmer & Lemeshow, 1989) as follows,

$$\Pr ob(U|X,G) = \frac{e^{(1-U)(-\beta_0-\beta_1 X-\beta_2 G-\beta_3 X*G)}}{1+e^{(1-U)(-\beta_0-\beta_1 X-\beta_2 G-\beta_3 X*G)}}. \qquad (28)$$

The usual likelihood ratio tests of significance of the estimates of the coefficients β_2 and β_3 address questions concerning uniform and nonuniform DIF, respectively. The difference in the log of the likelihood functions obtained in regressions with and without the β_3 coefficient is used to test for nonuniform DIF, or the significance of β_3. The difference in the log of the likelihood functions obtained in regressions with and without the β_2 coefficient is used to test for uniform DIF, or the significance of β_2.

The LR procedure requires a dichotomous dependent variable. To extend it to accommodate polytomous items, polytomous data must be recoded into a number of dichotomous sets, each of which is then suitable for a separate regression analysis. There are many extensions of the LR procedure to accommodate polytomous data (Agresti,

1990). French and Miller (1996) recommended three extensions of the logistic modeling for multinomial response models using ordinal response data, each of which involves a different set of pairwise comparisons between score categories or combinations of score categories. These models, called *continuation ratio logits*, *cumulative logits*, and *adjacent categories*, use the logit, or the ratio of the probability of getting the item (category) correct to the probability of getting the item (category) incorrect (Swaminathan & Rogers, 1990). For an item with K categories, each coding scheme produces $K - 1$ regressions.

The LR procedure can be used to identify both uniform and nonuniform DIF (French & Miller, 1996). However, using the LR procedure for DIF detection with polytomous data is quite unwieldy and time-consuming. Large amounts of data manipulation are required in applying LR to polytomous data; the dichotomization of the data through the three coding schemes reduces the power of the LR statistic because of the loss of data; therefore, large sample sizes are needed. Furthermore, it is awkward to use in an omnibus fashion. Interpretation of the results is difficult and it is difficult to come up with overall conclusions about the presence of DIF based on K - 1 separate regressions (French & Miller, 1996). Therefore, the LR technique is not included in this study.

Logistic Discriminant Function Analysis (LDFA)

Discriminant function analysis can be applied to a DIF study context to classify examinees into different groups which they most closely resemble on the basis of a set of item responses. Miller and Spray (1993) proposed to use the logistic discriminant

function analysis as a means of DIF identification of items that are polytomously scored. The *logistic discriminant function analysis* is more appropriate than normal discriminant function analysis in situations where the assumptions of the latter procedure, such as the normality of the explanatory variables, are often violated (Bull & Donner, 1987; Hosmer, Hosmer, & Fisher, 1983; Miller & Spray, 1993, Spray & Miller, 1994). In the logistic regression model (see Equation 28), the item response variable, U, is treated as a random variable and X and G are assumed to be fixed, explanatory variables. However, it has been shown that it is reasonable to use the logistic regression procedure to estimate Prob($G|X, U$) even though G is fixed and U is random (Anderson, 1972; Cox & Ferry, 1991; Hosmer & Lemeshow, 1989). In this form, Prob($G|X, U$) is simply a logistic form of the posterior probability used in discriminant analyses (Anderson, 1972). This procedure is called logistic discriminant function analysis (LDFA).

When applying the logistic discriminant function analysis to assess DIF in polytomous item responses, the discriminant function (without item notation) can be written as

$$\Pr ob(G|X,U) = \frac{e^{(1-G)(-\alpha_0-\alpha_1 X-\alpha_2 U-\alpha_3 XU)}}{1+e^{(1-G)(-\alpha_0-\alpha_1 X-\alpha_2 U-\alpha_3 XU)}}, \qquad (29)$$

The model is basically the same as Equation 28 except that the regression coefficients are now given by α_i, $i = 0, 1, 2, 3$ to distinguish them from the coefficients β_i in Equation 28 and the group variable, G ($G = 1$ for the reference group and $G = 0$ for the focal group), is still an indicator variable. The significance of Equation 29 is that, the response variable,

U, needs not be restricted to only two categories in this form but can take on item responses with any number of categories.

Tests of significance of the coefficients α_3 and α_2 provide answers to the questions concerning nonuniform and uniform DIF, respectively. The difference in the log of the likelihood functions obtained from Equation 29 and Equation 30 is used to test for nonuniform DIF, or the significance of α_3.

$$\Pr ob(G|X,U) = \frac{e^{(1-G)(-\alpha_0-\alpha_1 X-\alpha_2 U)}}{1+e^{(1-G)(-\alpha_0-\alpha_1 X-\alpha_2 U)}}. \tag{30}$$

Likewise, the difference in the log of the likelihood functions obtained from Equation 30 and Equation 31 is used to test for uniform DIF, or the significance of α_2.

$$\Pr ob(G|X,U) = \frac{e^{(1-G)(-\alpha_0-\alpha_1 X)}}{1+e^{(1-G)(-\alpha_0-\alpha_1 X)}}. \tag{31}$$

Equation 31 represents the discriminant function of null DIF in the item.

For each model, a likelihood ratio chi-square goodness-of-fit statistic, denoted G^2, can be computed. Differences in G^2 values between hierarchical models are asymptotically distributed as chi-square with one degree of freedom and provide a test of significance of the improvement in fit brought about by adding an additional term. The test statistic, denoted as G^2_{diff}, is given by

$$G^2_{diff} = -2(\ln L_{(i+1)} - \ln L_{(i)}), \tag{32}$$

where $L_{(i+1)}$ and $L_{(i)}$ are the likelihood's for adjacent hierarchical models. If the test of uniform or nonuniform DIF is significant, it means that for at least one of the item score categories, the probability of group membership, given the item score and the observed test score, differs significantly from that which would be predicted from the observed

score alone. Miller and Spray (1993) suggested that the estimated discriminant functions of the relevant item score categories can then be plotted in order to observe the direction of DIF.

Empirical Performance of the Polytomous DIF Detection Procedures

Zwick, Donoghue, and Grima (1993) conducted a simulation study to evaluate the Type I error and power of the Mantel (Mantel, 1963) and the generalized Mantel-Haenszel (GMH) procedures in detecting DIF with polytomous items following the partial credit model (Masters, 1982). They also examined the utility of $STND_{ES\text{-}DIF}$ (called SMD in their paper) in describing DIF. The factors that were varied across the simulation conditions were group ability distributions (2 levels: equal and unequal) and characteristics of the studied item (27 studied items: 4 patterns of DIF x 2 magnitudes of DIF x 3 sets of studied reference group parameters, plus, 3 null-DIF items). The four patterns of DIF were: a) DIF that was constant across response categories (constant DIF), b) DIF that changed signs across score categories (balanced DIF), c) DIF that affected only the lower score categories (low shift DIF), d) DIF that only affected the higher score categories (high shift DIF). The balanced DIF was a non-unidirectional DIF condition. The other three patterns of DIF were all unidirectional. Two magnitudes of DIF used were .1 and .25. Only a single test length was used (25 items). The matching test in this study consisted of 20 dichotomous and 4 polytomous items. Dichotomous items were generated using a three-parameter logistic model. Polytomous items were generalized using a partial credit model (Masters, 1982). They also examined the effect of different

ways of computing the matching variable on the performance of the Mantel and GMH procedures: whether or not the studied item was included in the matching variable (2 levels) and whether or not scores on polytomous items were rescaled in computing the matching variable (2 levels), i.e. whether to sum item scores or to assign different weights to dichotomous and polytomous items. For each condition, the Mantel and GMH procedures were performed and the means and standard deviations of $STND_{ES\text{-}DIF}$ were computed.

In terms of the matching variable, Zwick et al. (1993) showed that the studied item should be included in the matching variable and also scores on polytomous items should not be rescaled when calculating the matching variable. Excluding or rescaling the polytomous item when computing the matching variable resulted in increased Type I error. The results indicated that the appropriate method of forming the matching variable was to take the simple sum of all item scores, including the studied item. As for the performance of the Mantel and GMH procedures, the result showed that for the constant DIF conditions, the Mantel procedure was more powerful than the GMH. In the balanced DIF condition, the GMH was superior. For the high and low shift DIF conditions, the two procedures produced similar rejection rates. The choice of one over the other depends on the purpose of the research. When comparing group means is of interest, the Mantel (Mantel, 1963) procedure is a better choice. For most DIF analyses of polytomous items with ordered score categories, the Mantel approach would be more useful. If comparing groups in terms of their entire response distribution is of interest, the GMH should be used. The GMH could be particularly useful in analyzing unordered distractors, omits, or

solution strategies. As for the performance of the $STND_{ES\text{-}DIF}$, the results showed that the $STND_{ES\text{-}DIF}$ statistic was sensitive mainly to constant DIF. The mean $STND_{ES\text{-}DIF}$ was close to zero in the other DIF conditions.

Zwick, Thayer, and Mazzeo (1997) conducted a large-scale simulation study that evaluated the behaviour of three descriptive and five inferential procedures for assessing DIF in polytomous items. The descriptive statistics were $STND_{ES\text{-}DIF}$ (called SMD in their paper), standard SIBTEST DIF measure (with guessing correction), and modified SIBTEST DIF measure (without guessing correction). The two SIBTEST effect size measures were simply the numerators of the polytomous extension of the SIBTEST statistics. The five inferential procedures were: a test obtained by dividing $STND_{ES\text{-}DIF}$ by a standard error derived under the hypergeometric model (called SMD-H in their paper), a test obtained by dividing $STND_{ES\text{-}DIF}$ by a standard error derived under the multinomial model (called SMD-M in their paper), Mantel, Standard SIBTEST (with guessing correction), Modified SIBTEST (without guessing correction). The factors considered were focal group ability distributions (2 levels: $N(0,1)$ & $N(-1,1)$, the reference group ability was always $N(0,1)$), matching test length (2 levels: 20 and 50 items), and three properties of the studied item: discrimination (3 levels: .47, .86, 1.57), reference group item difficulty (2 levels: -.5, .5), and the difference between reference and focal group item difficulty (3 levels: .25, 0, -.25). Crossing all these factors yielded 72 conditions. Each DIF analysis was repeated 500 times, with samples of 500 examinees from each group. The results showed that, when the reference and focal groups had identical ability distributions, $STND_{ES\text{-}DIF}$ performed best as a descriptive statistic, but when the two

groups had unequal ability distributions, Modified SIBTEST tended to produce the best results. When the two groups had the same ability distribution, the five inferential procedures performed almost indistinguishably. When the ability distributions differed and the studied item discrimination was higher than the average discrimination of the matching items, the SIBTEST procedures showed much better Type I error control than Mantel and two test statistics based on $STND_{ES\text{-}DIF}$. The power ranking of the five inferential procedures depended on the direction of DIF and other factors. In general, Modified SIBTEST (without guessing correction) performed better than Standard SIBTEST (with guessing correction), both descriptively and inferentially.

Welch and Hoover (1993) compared the Mantel procedure with two combined *t* statistics: HW1 and HW3. HW1 tests a hypothesis about the difference between two mean values summed across *m*-independent *t* tests. HW3 uses a weighting procedure to more accurately represent the size of the focal and reference groups. Both HW1 and HW3 fall within the general standardization framework in which differences in expected item scores are averaged across levels of the matching variable using weights that are driven by statistical considerations (Potensa & Dorans, 1995). Welch and Hoover (1993) generated data using the partial credit model (Masters & Wright, 1984). All DIF items were of the constant type, i.e., DIF was constant across response categories. An external criterion (performance on a multiple-choice test) was used as the matching variable. Ten polytomous studied items were created by varying degrees of DIF ranging from 0 to .45. Four sample size ratios were crossed with two ability distributions (equal and unequal). The results indicated that the Mantel procedure was less powerful than HW1 and HW3

especially when differences in ability distribution existed. HW3 tended to be the most powerful in identifying DIF. However, the Type I error rates for HW1 and HW3 were more likely to exceed the nominal level than the error rate of the Mantel procedure, which displayed excellent Type I error control.

Chang, Mazzeo, and Roussos (1990) conducted two simulation studies to compare the Type I error and power of the Mantel procedure, a test statistic based on $STND_{ES\text{-}DIF}$ (called SMD in their paper), and SIBTEST. They employed a test statistic for $STND_{ES\text{-}DIF}$ using the standard error developed for the polytomous SIBTEST (Equation 27). In their first study, they generated data using the same specifications as those in Zwick et al. (1993). The results showed that all three procedures were useful in detecting unidirectional DIF items. The results for the Mantel procedure were very similar to those reported in Zwick et al. (1993). The results for SIBTEST were very similar to those obtained with the Mantel procedure, although the Mantel and the test statistic based on $STND_{ES\text{-}DIF}$ performed better than SIBTEST in terms of power and Type I error when the reference and focal groups had different ability distributions.

In their second study, Chang et al. (1990) used a wide range of item discrimination. Polytomous data were generated using Muraki's (1992) generalized partial credit model (GPCM) with varying discrimination parameters, 11 different values ranging from .15 to 2.0. A common set of difficulty parameters (-1, 0, 1) was used for all the studied items. One pattern of DIF (constant DIF), one magnitude of DIF (.25) and unequal group ability distribution (N(0,1) for reference group and N(-1,1) for focal group) were used. The major purpose was to examine the effect of studied item

discrimination parameters on the performance of the three procedures. The result of the study demonstrated the superiority of SIBTEST to the Mantel and the test statistic based on $STND_{ES\text{-}DIF}$ in controlling Type I error when the discrimination parameters of the studied items varied. Under certain conditions such as extremely high or low discriminations and very large sample size, the Mantel and the test statistic based on $STND_{ES\text{-}DIF}$ produced poor results in terms of power and Type I error. In contrast, SIBTEST showed only small changes in Type I error rates as a function of studied-item discrimination parameters.

Miller and Spray (1993) used real examinee data from an experimental mathematics performance test developed by ACT to demonstrate the utility of the LDFA for DIF identification of polytomously scored items. The test, administered to 1006 males and 971 females, consisted of 21 dichotomously scored multiple-choice and graded-response items and 6 polytomously scored open-ended items. The score categories for the open-ended items varied, ranging from 0-3 to 0-6. The LDFA procedure was applied to all 27 items, using the simple sum of item scores on all 27 items as the matching variable. For comparison purpose, the MH (Mantel & Haenszel, 1959) procedure and the Mantel (Mantel, 1963) were also applied to the dichotomous and polytomous data, respectively.

For dichotomously scored items, the results of the LDFA and the MH were similar in the detection of uniform DIF. Both identified 7 multiple-choice items as showing uniform DIF at $p<.001$. In addition, the LDFA procedure identified 2 multiple-choice items as showing nonuniform DIF at $p<.005$. For the open-ended items, both the LDFA and the Mantel procedures identified the same 2 items as showing uniform DIF. In

addition, the LDFA procedure identified 3 items as showing nonuniform DIF. To provide a means of interpreting the direction and practical significance of DIF for the statistically significant items, Miller and Spray (1993) also plotted the estimated discriminant functions for the null and full models along with a 95% confidence band around the estimated full model. Based on the inspection of the plots, DIF of any practical importance was identified only for one multiple-choice item and one open-ended item. Miller and Spray (1993) concluded that for purpose of detecting DIF with polytomous items, the LDFA procedure is superior to the polytomous logistic regression and Mantel procedures in that it is simpler and more practical than polytomous extensions of the logistic regression procedure and it is useful for identifying nonuniform DIF while the Mantel procedure is not.

Spray and Miller (1994) also did a simulation study to compare LDFA, Mantel, and GMH in detecting nonuniform DIF. Item responses to a twenty-item test were generated using the generalized partial credit model (Muraki, 1992). Only the last item had simulated DIF. Two sample sizes were used for each group: 500 and 2000. Abilities for both the reference and focal groups were sampled from a standard normal distribution. Three DIF conditions were simulated. Condition 1 was a uniform DIF condition where the *a*-parameters for both groups remained the same but the *b*-parameters were shifted by a constant amount. In Condition 2, nonuniform DIF was simulated where the *a*-parameters for each group varied but the *b*-parameters remained the same. In Condition 3, the *a*-parameter for each group remained the same and only two of the *b*-parameters varied, but in different directions. The item was easier for the reference group for the

second category but more difficult for the third category. The authors called it nonuniform DIF, but it really was a nonunidirectional uniform DIF condition (called balanced DIF in Zwick, et al., 1993). The results showed that the Type I error rates were within reasonable ranges of the nominal level for all three procedures, for both sample sizes, and under all DIF conditions. The LDFA procedure was powerful in identifying both uniform and nonuniform DIF simulated in all three conditions. Both Mantel and GMH were powerful in identifying uniform DIF simulated in Condition 1; however, neither identified nonuniform DIF simulated in Condition 2. The GMH procedure was more powerful than the LDFA procedure in detecting DIF simulated in Condition 3, indicating it was more sensitive to the directional change across response categories.

Relationships Among the Polytomous DIF Detection Procedures Studied

In the previous two sections, the main non-IRT procedures that have been proposed in the recent past to assess DIF with polytomous item responses were reviewed. SIBTEST is the only latent variable approach. All other procedures described above use an observed score measure of the construct of interest as a matching variable. A fundamental difference between the latent variable approaches and the observed score approaches is the use of estimates, derived from observed data, of the latent trait or true score instead of observed score as either an implicit or explicit matching variable (Potenza & Dorans, 1995). The STND, Mantel, and the GMH procedures are closely related to each other. Their null DIF definitions have been shown to be equivalent (Dorans & Holland, 1993; Potenza & Dorans, 1995; Zwick & Thayer, 1996).

Specifically, there is no DIF between groups after they have been matched on an observed score measure of ability. SIBTEST, whether dichotomous or polytomous version, assesses DIF in the same manner as the STND approach, with one important difference. Instead of using the empirical item-test regression used by STND, SIBTEST regresses item performance onto an estimate based on classical test theory of matching variable true score (Potenza & Dorans, 1995). The LDFA procedure is fundamentally different from the other five methods in that the group membership is an indicator instead of a predicator. The question addressed when using the LDFA procedure is whether the probability of group membership, given the item score and observed test score, differs significantly from that which would be predicted from the observed test score alone. In contrast, the question addressed using the other methods is whether, for persons of equal ability, predictions of item performance differ as a function of group membership (Miller & Spray, 1993).

The Mantel and STND procedures performed well when the data followed a partial credit model but yielded unacceptable results from the perspectives of both Type I error rate and power when the studied item discrimination parameters were extremely low or high and when two groups had unequal ability distributions. In contrast, SIBTEST is much more robust to violation of Rasch model conditions. The Mantel, the GMH, and STND procedures are easy to use and interpret, but they seemed not well suited for identifying nonuniform DIF. The LDFA procedure can also be used to identify nonuniform DIF. It is simpler and more practical than the LR procedure and it has

received the least attention. There is limited research in which the LDFA procedure was examined.

The Need to Assess the Polytomous DIF Detection Procedures

The results of the studies on the performance of the various polytomous DIF detection procedures are very encouraging. However, the conditions under which different procedures were examined are limited. In Welch and Hoover (1993) and Zwick et al. (1993), data were generated using the partial credit model with a fixed item discrimination parameter. Chang et al. (1996) varied the studied-item discrimination parameters but held the average discrimination of the matching test constant. Also, Chang et al. (1996) and Zwick et al. (1993) used equal sample sizes for the reference and focal groups, which is often not the case in practice. Another limitation is that in most studies reviewed in the previous section, a single matching test length was used. Zwick et al. (1997) investigated the performance of three descriptive and five inferential procedures by crossing factors such as ability distribution, matching test length, item discrimination and item difficulty. However, only one equal sample size was employed. The LDFA procedure (Miller & Spray, 1993) seems promising. However, it has not received an extensive simulation study in which its Type I error rate and power can be assessed.

Whereas DIF detection methodology is well-defined for traditional dichotomously scored multiple-choice items, there is not yet a consensus about the appropriate procedures that can be used to assess DIF in polytomously scored test items. Generally speaking, the statistical procedures described in the earlier sections have only recently

been proposed to study DIF for polytomous items. Limitations or complications exist in using each of them. Their utilities in assessing DIF in polytomous items have not received the comprehensive study accorded to their dichotomous DIF counterparts. The conditions examined are very limited. Further research is needed to investigate their Type I error and statistical power.

Research Questions

The purpose of this study was to investigate and compare the performance of DIF detection procedures in the identification of DIF with polytomous item responses under a variety of conditions. Five procedures were investigated in this study: Mantel, GMH, STND, SIBTEST, and LDFA. A simulation study was conducted to evaluate their performance when the properties of the items were known. As has been explained earlier, the Mantel approach is the extension of the MH procedure for ordered response categories and GMH procedure is the extension of the MH procedure for nominal data. Assessing DIF in polytomous items with ordered response categories is the primary concern of this study. Although the GMH method does not require that response categories be ordered, the focus of this study was on comparing it with other methods when ordering did, in fact, exist. With the LDFA procedure, the hypothesis of no uniform DIF and the hypothesis of no nonuniform DIF were tested separately using a likelihood ratio chi-square goodness of fit statistic. Therefore, the test statistic for both uniform and nonuniform DIF was included in this study, denoted as LDFA-Uniform and LDFA-

Nonuniform. The five DIF detection procedures were assessed with respect to their Type I error and power.

Researchers studying DIF detection procedures have indicated that the performance of most procedures is influenced by one or more factors such as sample size, test length, item parameters, matching criterion variable, group ability distribution, etc. In this study, the following variables were included to determine their effect on various polytomous DIF detection procedures: group ability distribution, test length, sample size, sample size ratio between focal and reference groups, studied-item discrimination parameter, and type of DIF.

Specifically, the following research questions are addressed:

1. What are the Type I error rates associated with the Mantel, GMH, STND, SIBTEST, and LDFA procedures when they are used to detect DIF in polytomous item responses simulated with no DIF?

2. How powerful are the Mantel, GMH, STND, SIBTEST, and LDFA procedures when they are used to detect DIF in polytomous item responses simulated with DIF?

3. Do group ability distribution, test length, sample size, sample size ratio between focal and reference groups, discrimination parameter of the studied item, and type of DIF have any effect on the Type I error and power of the Mantel, GMH, STND, SIBTEST, and LDFA procedures?

CHAPTER III
METHODOLOGY

The methods used to assess the Type I error and power of the Mantel, GMH, STND, SIBTEST, and LDFA DIF detection procedures are presented in this chapter. The chapter is divided into six sections. First, the conditions under which these statistical procedures were compared are described in detail. They include factors held constant and factors manipulated. Next, the item response models used in this study are presented. Then, item parameters used to simulate item responses are identified. Next, the steps used in generating data are outlined. This is followed by a description of the software used to compute each DIF statistic. Finally, the chapter ends with the proposed data analysis method.

Simulation Conditions

In this section, the conditions under which the polytomous DIF detection statistics are computed are elaborated. Polytomous item response data were simulated under a variety of conditions, with each data set accommodating prespecified levels of a number of different factors that may have an effect on the Type I error and power of the Mantel, GMH, STND, SIBTEST, and LDFA procedures.

To evaluate the Type I error of each DIF detection procedure, items with no DIF were simulated. The factors manipulated were: group ability distribution, test length, sample size, sample size ratio between focal and reference groups, and studied-item discrimination parameter.

To evaluate the power of each DIF detection procedures, items with DIF were simulated. The factors included were: group ability distribution, test length, sample size, sample size ratio between focal and reference groups, studied-item discrimination parameter, and DIF condition.

Other factors such as the number of groups, the number of studied items, and matching variable were held constant for all the conditions examined and for both Type I error and power studies. In addition, the magnitude of DIF was held constant for the power study.

Factors Held Constant

Number of Groups. The DIF statistics were compared for conditions where the number of groups is equal to two, i.e., the reference group and the focal group. The group suspected of being disadvantaged, usually the minority group, is called the focal group, and the group against which the performance of the focal group is compared, usually the majority group, is called the reference group.

Number of Studied Items. Only one studied item was examined in each condition. All the studied items simulated in this study were polytomous items with four response categories (scored 0, 1, 2, 3).

DIF Magnitude. DIF magnitude is the amount of DIF an item shows. One DIF magnitude of .25 was used to simulate items with uniform DIF in the power study. A DIF magnitude of .25 is often common in actual testing situations (Zwick, et al., 1993; Chang, et al., 1996; Spray & Miller, 1994).

Matching Variable. The comparison of matched or comparable groups is critical in a DIF study because it is important to distinguish between differences in item functioning and differences in group ability levels. In addition, all of the methods in this study were conditional on a measure of ability. Conditioning was needed because it is important to dissociate impact from DIF in any DIF study (Holland & Thayer, 1986). In this study, both the matching variable and the conditioning variable were the simple sum of the item scores. Zwick et al. (1993) showed that the studied item should be included in the computation of the matching variable for the Mantel, GMH, and STND (called SMD in their study) in order to produce unbiased estimates. In this study, to control the impact-induced Type I error inflation, the matching variable was computed by including the studied item for the Mantel, GMH, and STND statistics. Similarly, LDFA uses the simple sum of item scores as the conditioning variable and SIBTEST uses the regression correction to correct for Type I error inflation due to impact.

Factors Manipulated

Ability Distribution. It is important to assess the effectiveness of DIF statistics when the groups differ in their ability distributions. Research has shown that most DIF detection procedures worked well when comparison groups had equal ability distribution, but displayed poor Type I error performance when groups of differing abilities were compared (e.g. Zwick et al., 1993). The impact of the difference in underlying ability distributions was investigated by examining three different conditions. In the first condition, the reference and focal groups had identical ability distributions. That is, the ability parameter for each group was randomly selected from a standard normal $N(0,1)$

distribution. In the second condition, while the reference group ability was still sampled from a $N(0,1)$ distribution, the focal group ability was sampled from a $N(-.5,1)$ distribution resulting in a lower ability level than that of the reference group. Ability distribution that differed by .5 standard deviation simulated the case in which there is not a very substantial between-group difference. In the third condition, while the reference group ability was still sampled from a $N(0,1)$ distribution, the focal group ability was sampled from a $N(-1,1)$ distribution. This condition simulated the case in which there is a substantial between-group difference. A difference in group ability means of .5 and 1 standard deviation is often common in actual testing situations.

Test Length. Two test lengths were used: twenty items and forty items. Both tests included eighty percent dichotomous items and twenty percent polytomous items. That is, the twenty-item test included 16 dichotomous items and 4 polytomous items, and the forty-item test included 32 dichotomous items and 8 polytomous items. The last item in each test (i.e., Item 20 in the twenty-item test and Item 40 in the forty-item test) was used as the studied item. All the studied items simulated in this study were polytomous items with four response categories (scored 0, 1, 2, 3). Longer tests tend to be more reliable than the shorter tests. Therefore, any distortions in the functioning of the statistical DIF detection procedures were expected to be less severe with the forty-item test than with the twenty-item test.

Sample Size and Sample Size Ratio. Sample size and sample size ratio between the focal and reference groups may affect the robustness of the statistics through their effect on estimation (Swaminathan & Rogers, 1990). Research conducted on the

performance of the various dichotomous DIF statistics indicated that power increases with increased sample size. Also both equal and unequal sample sizes between the reference and focal groups may be obtained in DIF analyses in real testing situations, depending on which subgroups are of interest. For example, when gender-based DIF is a major concern, approximately equal sample sizes can often be obtained for males and females. However, when the comparison between Caucasian and African American examinees is of primary interest, the group size of African American examinees is often much smaller. Therefore, both equal and unequal sample sizes between reference and focal groups were used in this study. Sample size ratios between focal and reference groups were set at: 1:1 and 1:2. For each ratio, three sample sizes were used, representing relatively small, medium, and large sample sizes. The total numbers of examinees including both the focal and reference groups were 600, 1200, and 2400 for the small, medium, and large sample-size representations, respectively and they were fixed to be the same for both the 1:1 and 1:2 ratios Thus, six sample size combinations between focal and reference groups were simulated. They were 300/300, 600/600, 1200/1200, 200/400, 400/800, 800/1600 for the focal and reference groups, respectively. These sample sizes were chosen to represent actual DIF study conditions.

Item Discrimination. To study the Type I error and power of the polytomous DIF procedures as a function of studied-item discrimination parameter, three levels of *a*-parameters were used: 0.5, 1.0, and 1.5 for the null and uniform DIF items, each representing relatively low, medium, and high discrimination respectively; and three levels of *a*-parameter difference between the focal and reference groups were used: 0.5,

1.0, and 1.5 for the nonuniform DIF items. Previous theoretical and simulation work demonstrated the relevance of this factor to DIF (Chang, et al., 1996; Zwick, et al., 1997). The amount of DIF in the studied item as measured in terms of differences between item response functions depends on both the threshold and the discrimination parameter values (Chang et al., 1996).

DIF Condition. The first 19 items in the twenty-item test and the first 39 items in the forty-item test were free of DIF. The last item in each test (i.e., Item 20 in the twenty-item test and Item 40 in the forty-item test) played the role of the studied item. To evaluate the Type I error rate of each DIF detection procedure, items with no DIF were generated. To evaluate the power of each DIF detection procedure, items with DIF were simulated. The studied item always had four response categories. Four DIF conditions were examined.

The first condition was the *null-DIF condition* in which the reference and focal groups had identical studied-item parameters. This condition was designed for the Type I error study. The other conditions described below were designed for the power study.

The second condition was a simple unidirectional uniform-DIF case where the *a*-parameters for both groups remained the same but the *b*-parameter was shifted by a constant amount (.25) across score categories. This condition was referred to as the *uniform-DIF (constant) condition* in this study. All of the transitions from a given score category to the next highest category were set to be more difficult for members of the focal group than for comparable members of the reference group.

The third condition was a balanced uniform-DIF case where the *a*-parameters for each group once again remained the same and only two of the *b*-parameters varied, but in different directions. The transition from the lowest to the second category was more difficult for the focal group ($b_{j1F}=b_{j1R}+.25$). The transition from the second to the third category was the same for the two groups ($b_{j2F}=b_{j2R}$). The transition from the third category to the fourth category was easier for the focal group ($b_{j1F}=b_{j1R}-.25$). This condition was referred to as the *uniform-DIF (balanced) condition* in this study. This type of DIF has been investigated in several studies under different research conditions (e.g., Allen & Donoghue, 1994; Chang, et al., 1996; Spray & Miller, 1994; Zwick, et al., 1993)

The fourth condition was a *nonuniform-DIF condition* where the *a*-parameters differed but the *b*-parameters remained the same for the focal and reference groups.

Twelve studied items were simulated, three for each of the four DIF conditions. The parameters for the twelve studied items are given later in this chapter.

Final Design. In summary, the factors considered in the simulation were group ability distribution differences (three levels), test length (two levels), sample size (three levels), sample size ratio (two levels), item discrimination (Null & Uniform DIF) or item discrimination parameter difference between reference and focal groups (Nonuniform DIF) (three levels), and DIF conditions (four levels). Table 2 summarizes the factors and their levels used in this study. Data sets in each of the 3x2x3x2x3x4 conditions (ability distribution differences by test length by sample size by sample size ratio by item discrimination parameter or parameter difference between R and F groups by DIF

Table 2
Simulation Design

Ability Distribution Differences: 3 levels

1. equal: both the reference and focal group ability distributions are standard normal ($N(0,1)$)
2. unequal: the ability distribution of the reference group is standard normal ($N(0,1)$), and the ability distribution of the focal group is ($N(-.5,1)$)
3. unequal: the ability distribution of the reference group is standard normal ($N(0,1)$), and the ability distribution of the focal group is ($N(-1,1)$)

Test Length: 2 levels

1. 20 items: 16 dichotomous items and 4 polytomous items
2. 40 items: 32 dichotomous items and 8 polytomous items

Sample Size Combination: 6 combinations (2 ratios x 3 sample sizes (small, medium, large))

1. the sample size ratio between the focal and reference groups is 1:1: 300/300, 600/600, 1200/1200
2. the sample size ratio between the focal and reference groups is 1:2: 200/400, 400/800, 800/1600

Studied-Item Discrimination Parameter (Null and Uniform DIF Items) or Discrimination Parameter Difference Between R and F Groups (Nonuniform DIF Items): 3 levels each

Null & Uniform DIF	Nonuniform DIF
1. a=0.5	1. a_{jR}-a_{jF}=0.5
2. a=1.0	2. a_{jR}-a_{jF}=1.0
3. a=1.5	3. a_{jR}-a_{jF}=1.5

DIF Conditions: 4 levels

1. null-DIF: 3 items (a_{jF}=a_{jR}=0.5 or 1.0 or 1.5, b_{jkF}=b_{jkR}=(-1,0,1), k=1,2,3)
2. uniform DIF (constant): 3 items (a_{jF}=a_{jR}, b_{jkF}=b_{jkR} +.25)
3. uniform DIF (balanced): 3 items (a_{jF}=a_{jR}, b_{j1F}=b_{j1R} +.25, b_{j2F}=b_{j2R}, b_{j3F}=b_{j3R} -.25)
4. nonuniform DIF: 3 items (b_{jR}=b_{jF}, a_{jF}=.5, a_{jR}=a_{jF}+.5, or 1.0, or 1.5)

Total Simulation Conditions: 432 (3 ability distribution differences x 2 test lengths x 6 sample size combinations x 3 a-parameters or a-parameter differences x 4 DIF conditions). For each condition, 100 replications were carried out.

condition) were replicated 100 times. This yielded a total of 43,200 data sets. Each data set was, in turn, analyzed with the each DIF detection procedure (Mantel, GMH, STND, SIBTEST, & LDFA).

Item Response Models Used in This Study

For dichotomous multiple-choice items, the three-parameter logistic (3PL) model described below is used to simulate item responses. For a randomly chosen examinee with ability θ , the probability of answering item *j* correctly, $P_j(\theta)$, is given by

$$P_j(\theta) = c_j + \frac{1 - c_j}{1 + \exp\left\{-1.7a_j(\theta - b_j)\right\}}, \tag{33}$$

where a_j, b_j, and c_j represent item discrimination, difficulty, and pseudo-guessing parameters, respectively. Specific item parameter values used to simulate data are identified in the next section.

Polytomous item responses were simulated using Muraki's (1992) generalized partial credit model (GPCM), which gives the item category characteristic curves (ICCCs) as functions of an unidimensional latent ability, θ. Assume that an item is scored on a scale with $m+1$ ordered categories ranging from *0* to *m*. With the GPCM model, the probability of selecting the *k*th category from *m* possible categories of item *j* for an examinee with ability θ is given by

$$P_{jk}\left\{X_j = k|\theta\right\} = \frac{\exp\left\{.7a_j\sum_{v=0}^{k}(\theta - b_{jv})\right\}}{\sum_{c=0}^{m}\exp\left\{.7a_j\sum_{v=0}^{c}(\theta - b_{jv})\right\}}. \tag{34}$$

where b_{jv} is the threshold parameter. It represents the difficulty of making the transition from category m - 1 to category m and it defines the point of intersection of the adjoining ICCCs and a_j represents a slope parameter relating to the discriminating power of the item. By convention, $\sum_{v=0}^{0} a_j(\theta - b_{jv}) \equiv 0$. The GPCM model is analogous to a two-parameter logistic response model for ordered response categories.

Item Parameters Used in Simulation

Parameters for the first 19 items of the 20-item test are listed in Table 3 and parameters for the first 39 items of the 40-item test are listed in Table 4. These items had identical item parameters for the reference and focal groups. Item parameters for multiple-choice items were selected from an administration of the Graduate Management Admissions Test reported in Narayanan and Swaminathan (1994). The parameters for polytomous items were selected from an actual calibration for the 1992 NAEP (Johnson & Carlson, 1994). Some of these parameters were also used in Chang et al. (1996).

Parameters for the studied items are listed in Table 5. Altogether, twelve studied items were simulated, three for each DIF condition. A common set of reference group threshold parameters (-1, 0, 1) was used for all the studied items.

For the null DIF condition, studied items had identical item parameters. Three null-DIF items were simulated by combining one set of *b*-parameters (-1, 0, 1) with each of the three *a*-parameters (0.5, 1.0, and 1.5) (see Table 5).

For the uniform-DIF (constant) condition, the *a*-parameters for both groups remained the same but the *b*-parameters were shifted by a constant amount across score

Table 3
Item Parameters for Items 1-19
(20-item test)

Dichotomous Item Parameters			
Item Number	a_j	b_j	c_j
1	0.44	-0.30	0.20
2	0.82	1.02	0.20
3	1.02	1.28	0.20
4	0.92	0.42	0.20
5	0.56	-2.70	0.20
6	0.35	-1.12	0.20
7	1.05	0.10	0.20
8	0.73	0.61	0.20
9	1.11	-0.35	0.20
10	0.55	1.09	0.20
11	0.92	1.13	0.20
12	1.01	0.81	0.20
13	0.70	1.05	0.20
14	0.48	2.12	0.20
15	0.53	0.87	0.20
16	1.12	-1.21	0.20

Polytomous Item Parameters				
Item Number	a_j	b_{j1}	b_{j2}	b_{j3}
17	0.563	-2.449	-0.089	2.416
18	1.359	-0.342	1.008	1.797
19	0.779	-0.345	2.428	2.822

Table 4: Item Parameters for Items 1-39
(40-item test)

Dichotomous Item Parameters			
Item Number	a_j	b_j	c_j
1	0.44	-0.30	0.20
2	0.55	-1.06	0.20
3	0.82	1.02	0.20
4	0.52	-1.96	0.20
5	1.02	1.28	0.20
6	0.82	0.61	0.20
7	0.92	0.42	0.20
8	0.65	1.68	0.20
9	0.56	-2.70	0.20
10	0.29	-1.39	0.20
11	0.35	-1.12	0.20
12	0.31	-1.37	0.20
13	1.05	0.10	0.20
14	0.51	-0.09	0.20
15	0.73	0.61	0.20
16	0.88	0.95	0.20
17	1.11	-0.35	0.20
18	1.32	0.57	0.20
19	0.55	1.09	0.20
20	1.40	1.64	0.20
21	0.92	1.13	0.20
22	0.64	-1.55	0.20
23	1.01	0.81	0.20
24	0.61	-0.53	0.20
25	0.70	1.05	0.20
26	1.02	0.64	0.20
27	0.48	2.12	0.20
28	1.01	0.91	0.20
29	0.53	0.87	0.20
30	0.36	-2.63	0.20
31	1.12	-1.21	0.20
32	0.86	-0.57	0.20

Polytomous Item Parameters				
Item Number	a_j	b_{j1}	b_{j2}	b_{j3}
33	0.563	-2.449	-0.089	2.416
34	1.359	-0.342	1.008	1.797
35	0.535	-1.804	-0.368	0.219
36	0.779	-0.345	2.428	2.822
37	0.765	0.979	-0.719	-0.260
38	0.594	-0.535	-0.177	0.713
39	0.536	0.257	-1.083	0.826

Table 5
Parameters for the Studied Items

Null DIF

1. $b_{jkF}=b_{jkR}=(-1, 0, 1)$, $a_{jF}=a_{jR}=0.5$
2. $b_{jkF}=b_{jkR}=(-1, 0, 1)$, $a_{jF}=a_{jR}=1.0$
3. $b_{jkF}=b_{jkR}=(-1, 0, 1)$, $a_{jF}=a_{jR}=1.5$

Uniform DIF (constant)

4. $b_{jkR}=(-1, 0, 1)$, $b_{jkF}=(-0.75, 0.25, 1.25)$, $a_{jF}=a_{jR}=0.5$
5. $b_{jkR}=(-1, 0, 1)$, $b_{jkF}=(-0.75, 0.25, 1.25)$, $a_{jF}=a_{jR}=1.0$
6. $b_{jkR}=(-1, 0, 1)$, $b_{jkF}=(-0.75, 0.25, 1.25)$, $a_{jF}=a_{jR}=1.5$

Uniform DIF (balanced)

7. $b_{jkR}=(-1, 0, 1)$, $b_{jkF}=(-0.75, 0, 0.75)$, $a_{jF}=a_{jR}=0.5$
8. $b_{jkR}=(-1, 0, 1)$, $b_{jkF}=(-0.75, 0, 0.75)$, $a_{jF}=a_{jR}=1.0$
9. $b_{jkR}=(-1, 0, 1)$, $b_{jkF}=(-0.75, 0, 0.75)$, $a_{jF}=a_{jR}=1.5$

Nonuniform DIF

10. $b_{jkF}=b_{jkR}=(-1, 0, 1)$, $a_{jF}=0.5$, $a_{jR}=1.0$
11. $b_{jkF}=b_{jkR}=(-1, 0, 1)$, $a_{jF}=0.5$, $a_{jR}=1.5$
12. $b_{jkF}=b_{jkR}=(-1, 0, 1)$, $a_{jF}=0.5$, $a_{jR}=2.0$

categories. All of the transitions from a given score category to the next higher category were assumed to be more difficult for the members of the focal group than for comparable members of the reference group. The level of threshold shift was 0.25. Thus, the focal group thresholds for the studied items were (-0.75, 0.25, 1.25), compared to (-1, 0, 1) for the reference group. The threshold shift of 0.25 was paired with each of the three *a*-parameters (0.5, 1.0, and 1.5) to create three uniform-DIF (constant) items (see Table 5).

For the uniform-DIF (balanced) condition, the *a*-parameters for each group once again remained the same and only two of the *b*-parameters varied, but in different directions. Again, the level of threshold shift was 0.25. Thus, the focal group thresholds for the studied items were (-0.75, 0, 0.75), compared to (-1, 0, 1) for the reference group. The threshold shift of 0.25 was paired with each of the three *a*-parameters (.5, 1.0, and 1.5) to create three uniform-DIF (balanced) items (see Table 5).

Three nonuniform-DIF items were simulated by combining one set of *b*-parameters with each of the three *a*-parameter differences. The *b*-parameters for both groups were (-1, 0, 1) for all three studied items. The *a*-parameters for the focal group was 0.5 for all three studied items. The differences in *a*-parameters between the focal and reference groups were 0.5, 1.0, and 1.5. Thus, the *a*-parameters for the reference group were 1.0, 1.5, 2.0 respectively, for the three studied items. All the three nonuniform-DIF items were more discriminating for the reference group for all response categories (See Table 5).

Simulation Procedure

In this section, the steps which were used to simulate the data and to perform DIF detection procedures are presented. There were a total of 432 cells in the design. For each cell, 100 replications were carried out. Each replication involved the following steps:

Step 1

Generate θ values for the reference and focal groups, respectively, according to the appropriate ability distributions and sample size combinations.

Step 2

Generate item responses according to appropriate item response models and appropriate item parameters for each sampled θ value. The 3PL model was used to simulate dichotomous item responses and the GPCM model was used to simulate polytomous item responses.

Step 3

Perform the Mantel, GMH, STND, SIBTEST, and LDFA DIF detection procedures on the simulated item responses and record the resulting statistics.

Step 4

Steps 1 through 3 were repeated 100 times for each simulation condition.

Computer Programs

The program used to generate the dichotomous item responses was written using C language. The program used to generate polytomous item responses was originally written in Fortran by Spray (1994) and was modified and converted to C by Gong (1998). The Mantel statistic was computed using PSYCH.FOR (Spray, 1994; modified and converted to C by Gong, 1998). The GMH statistic was computed using GENMH.FOR (Spray, 1994; modified and converted to C by Gong, 1998). The program used to compute the STND statistic was written by Gong (1998) using C language. The SIBTEST statistic was computed using the software PolySIBTEST (Roussos, Chang, & Stout, 1996). The LDFA statistic was also computed using program PSYCH.FOR (Spray, 1994; modified and converted to C by Gong, 1998). The parameter estimates for the

LDFA were obtained using the Newton-Raphson maximum likelihood estimation procedure. Overall model fit was not tested in this simulation; only differences between models were tested.

Data Analysis

To assess the Type I error performance of each DIF detection procedure, the number of times a null-DIF item was falsely rejected at the .05 significance level, in each simulation condition, over 100 replications, was recorded. To assess the power of a DIF detection procedure, the number of times a DIF item was correctly identified at the .05 significance level, in each simulation condition, over 100 replications, was recorded. As mentioned in Chapter II, LDFA tests separately the hypothesis of no uniform DIF and the hypothesis of no nonuniform DIF; therefore, both LDFA-Uniform and LDFA-Nonuniform test statistics were reported and analyzed.

The effects of the manipulated factors on Type I error rates and power of DIF detection procedures were analyzed using logit analysis. A logit analysis is a log-linear analysis with one variable specified as the dependent variable. The main effect of any independent variable, with this analysis, actually represents an interaction with the dependent variable (Baker, 1981). Bock's approach (Bock, 1975, Chap.8) was followed to obtain the most parsimonious logit model. This approach used a nested hierarchical model, meaning that all the lower-order effects nested within the model were also included. The analysis was done in a forward stepwise manner. It starts with the simplest main effects and then fits incrementally more complex models until a hierarchical model

was found to be insignificant. The overall evaluation of the model was made using the likelihood ratio chi-square statistic. A model with a p-value greater than or equal to 0.15 was considered adequate. Any individual effect was considered to be significant if the size of absolute value of the standardized parameter estimate was greater than 1.96.

In this study, the outcome (i.e. reject or do not reject at α=.05) was used as the dependent variable for the logit analysis. The independent variables were test length, sample size, sample size ratio between the focal and reference groups, and studied-item *a*-parameter (null, constant uniform DIF, balanced uniform DIF) or *a*-parameter difference (nonuniform DIF). The original plan was to include group ability distribution difference as an independent variable. However, after some preliminary study, it was found that group ability distribution difference was a major factor affecting the performance of each procedure; all procedures seemed to perform well when there was no group ability difference but not so well when group difference existed; therefore it was necessary to analyze the results by the level of group ability difference. The logit analysis determined which independent variable or which combination of variables was significantly associated with rejection rates made by each DIF detection procedure under each level of group ability difference. The effects of the independent variables on the outcome variable were analyzed separately for each of the four DIF conditions and for each of the DIF detection procedures under investigation, namely, the Mantel, GMH, STND, SIBTEST, and LDFA (LDFA-Uniform, LDFA-Nonuniform) procedures.

CHAPTER IV
RESULTS

In this chapter, the results of this investigation are presented separately for each procedure. For the null DIF condition, because no group difference was introduced into the studied items, the observed rejection rates for this condition correspond to the Type I errors rates. In this study, the conventional significance level (α) of .05 was adopted as the nominal Type I error; thus, a Type I error within or at .05 is considered to be in good control. A Type I error above .05 but under .10 is considered to be inflated but still acceptable. A Type I error over .10 is considered to be excessive and unacceptable. For uniform-DIF (constant) and uniform-DIF (balanced) conditions, a group difference in item difficulty was introduced into the studied items; for nonuniform DIF, a group difference in item discrimination was simulated into the studied items; therefore, the empirical rejection rates depict the power of each procedure. In this study, statistical power below .70 was considered inadequate, power over .70 was considered to be adequate, and power over .80 was considered to be high. These criteria for Type I error and power were somewhat arbitrary but reasonable. They were established for the clarity and convenience of analysis and discussion. Since the power of a statistical procedure is related to its Type I error, the results for the power of each procedure are presented in conjunction with their Type I error.

The Type I error and power of all procedures were affected by group ability distribution difference in a similar way (except LDFA-Nonuniform). In general, the Type I error and power increased as the between-group ability difference increased. Group

ability distribution difference seemed to be a major factor affecting the performance of all procedures under investigation; they all performed well when there was no group ability difference but not so well when a group ability difference existed. Therefore, the results of each procedure were analyzed for each level of group ability difference separately.

Logit analysis was conducted for each procedure, for the appropriate DIF condition, and for each level of group ability difference separately in order to determine which factors had the most impact on their Type I error and power. The independent variables were test length (L), sample size (S), sample size ratio (R), and studied-item discrimination (A) or discrimination-parameter difference (A_D). The outcome (reject or do not reject at α=.05) was the dependent variable. As mentioned in Chapter III, the overall evaluation of the model was made using the likelihood ratio chi-square statistic. A model with a p-value greater than or equal to 0.15 was considered adequate. Any individual effect was considered to be significant if the size of the absolute value of the standardized parameter estimate was greater than 1.96. Logit analysis is a log-linear analysis with one variable specified as the dependent variable. The main effect of any independent variable, with this analysis, actually represents an interaction with the dependent variable. For simplicity, the results are presented in terms of the independent variables only. However, in logit analysis this implies an interaction with the dependent variable (the outcome of reject or do not reject at α=.05). For example, if a test length-by-outcome effect was significant, it was referred to only as a test length main effect.

The Mantel Procedure

Uniform DIF (Constant)

Equal Group Ability Distributions **(R-*N*(0,1), F-*N*(0,1))**

For Type I error, the result of logit analysis indicated that a model with only the dependent variable (i.e., the outcome of reject or do not reject at α=.05) was sufficient to explain the distribution of frequencies observed for Mantel (χ^2=14.19, df=35, p=.999). Therefore, none of the independent variables (i.e., L, A, S, and R) had any effect on the Type I error rates of Mantel.

For power, the result of logit analysis indicated that a model with the effects of A, S, R, and AxS was needed to sufficiently explain the frequencies of the outcome variable for Mantel (χ^2=23.61, df=18, p=.17). The detection rates for the AxS interaction for Mantel under the condition of equal group abilities are presented in Table 6. As can be seen from the table, the effect of item discrimination depends on the size of the sample; the difference among the three levels of item discrimination decreased as sample size

Table 6: Detection Rates for the Item Discrimination by Sample Size Interaction for the Uniform-DIF (Constant) Condition for the Mantel Procedure (R-*N*(0,1), F-*N*(0,1))

	Sample Size		
Item Discrimination	300/300 200/400	600/600 400/800	1200/1200 800/1600
a=0.5	.530	.768	.968
a=1.0	.753	.985	1.00
a=1.5	.965	.998	1.00

Note. Each entry is averaged over 2 test lengths and 2 sample size ratios and is based on 400 observations.

increased. This interaction can be explained by the ceiling effect observed as item discrimination and sample size increased. With respect to the main effects of A, S, and R, in general, the power of Mantel increased as sample size and item discrimination increased and decreased as sample size ratio changed from 1:1 to 1:2. Although included in the logit model, the effect of sample size ratio was trivial.

The Type I error rates and power (α=.05) of the Mantel procedure in detecting uniform DIF (constant) for the condition of equal group abilities are reported in Table 7. As can be seen from Table 7, the Type I error rates of Mantel were within or close to the expected theoretical value of .05 in most conditions. Some conditions produced slightly inflated Type I error rates. However, there were no clear patterns.

Table 7: The Type I Error and Power of the Mantel Procedure in Detecting Uniform DIF (Constant) for the Condition of Equal Group Ability Distributions (R-$N(0,1)$, F-$N(0,1)$)

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.03	0.05	0.03		0.57	0.73	0.95
	600/600		0.05	0.06	0.06		0.75	0.97	0.99
	1200/1200		0.07	0.05	0.04		0.96	1.00	1.00
1:2 Ratio	200/400		0.03	0.04	0.06		0.44	0.74	0.95
	400/800		0.05	0.06	0.04		0.74	0.98	1.00
	800/1600		0.06	0.04	0.07		0.97	1.00	1.00
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.05	0.07	0.04		0.59	0.84	0.98
	600/600		0.04	0.04	0.05		0.86	1.00	1.00
	1200/1200		0.03	0.07	0.03		1.00	1.00	1.00
1:2 Ratio	200/400		0.06	0.07	0.06		0.51	0.70	0.97
	400/800		0.03	0.04	0.06		0.72	0.99	1.00
	800/1600		0.06	0.07	0.04		0.94	1.00	1.00

Since Mantel had very good adherence to the nominal Type I error rate when there was no group ability difference, its power can be best assessed under this condition. One can only have great confidence in the observed power of a statistical test when its Type I error is under control. As has been explained at the beginning of this chapter, statistical power over .70 was considered to be adequate and power over .80 was considered to be high in this study. Given these criteria, Mantel was extremely powerful in detecting constant uniform DIF with the combination of medium (600/600, 400/800) to large sample size (1200/1200, 800/1600) and medium to high discrimination (a=1.0 or 1.5), the combination of small sample size (300/300, 200/400) and high discrimination (a=1.5), and the combination of large sample size and low discrimination (a=0.5). Under these conditions, the power ranged from .94 to 1.0. The power of Mantel was adequate in other conditions except with the combination of low discrimination and small sample size (with power ranging from .44 to .59). Sample sizes of 300/300 and 200/400 were too small for Mantel to have adequate power when item discrimination was low.

Unequal Group Ability Distributions **(R-*N*(0,1), F-*N*(-.5,1))**

The Type I error rates and power (α=.05) of the Mantel procedure in detecting uniform DIF (constant) for the condition of small group ability difference are reported in Table 8. The Type I error rates of Mantel increased considerably. In general, Type I error was in good control when item discrimination was low (a=0.5) except when sample size was large and sample size ratio was 1:1; Type I error was within the acceptable range, <.10, as set in this study when item discrimination was medium (a=1.0) except when sample size was large and test length was short. Most of the Type I error rates at a=1.5

Table 8: The Type I Error and Power of the Mantel Procedure in Detecting Uniform DIF (Constant) for the Condition of Unequal Group Ability Distributions (R-$N(0,1)$, F-$N(-.5,1)$)

			Type I Error				Power		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.02	0.04	**0.19**		0.50	0.92	**0.99**
	600/600		0.02	0.06	**0.20**		0.85	0.97	**1.00**
	1200/1200		0.08	0.18	**0.49**		0.95	1.00	**1.00**
1:2 Ratio	200/400		0.04	0.09	**0.15**		0.33	0.84	**0.97**
	400/800		0.04	0.07	**0.27**		0.74	0.97	**1.00**
	800/1600		0.06	0.14	**0.49**		0.95	1.00	**1.00**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.03	0.06	**0.13**		0.55	0.86	**0.98**
	600/600		0.05	0.08	**0.09**		0.80	1.00	**1.00**
	1200/1200		0.09	0.08	**0.17**		0.98	1.00	**1.00**
1:2 Ratio	200/400		0.03	0.05	**0.08**		0.41	0.88	**0.97**
	400/800		0.04	0.08	**0.14**		0.65	1.00	**1.00**
	800/1600		0.05	0.08	**0.11**		0.95	1.00	**1.00**

exceeded the acceptable rate (>.10) as set in this study. This was especially true for the 20-item test. Consequently, the power under this condition was not credible. Given the unacceptably high Type I error rates for high discrimination, it is not meaningful to further assess the effects of other factors. Therefore, only two levels of item discrimination (a=0.5 and a=1.0) were used in the logit analysis for the condition of small group ability difference. The column under a=1.5 in Table 8 was highlighted and was not included in the logit analysis.

The result of the logit analysis indicated that for Type I error, a model with the main effects of A and S was adequate to explain the frequency distribution of the dependent variable for Mantel (χ^2=13.37, df=20, p=.861). Increasing sample size and item discrimination led to an increase in Type I error rates.

For power, a model with the effects of A, S, R, and AxS was needed to explain the frequencies of the outcome variable for Mantel (χ^2= 19.04, df=17, p=.326). The detection rates for the AxS interaction for Mantel under the condition of small group ability difference are presented in Table 9. As can be seen from the table, the effect of item discrimination depends on the size of the sample; the difference between the two levels of item discrimination decreased as sample size increased. This was due to the ceiling effect observed at a=1.0. Also shown in Table 9, sample sizes of 300/300 and 200/400 were too small for Mantel to have adequate power when item discrimination was low.

Table 9: Detection Rates for the Item Discrimination by Sample Size Interaction for the Uniform-DIF (Constant) Condition for the Mantel Procedure (R-N(0,1), F-N(-.5,1))

	Sample Size		
Item Discrimination	300/300 200/400	600/600 400/800	1200/1200 800/1600
a=0.5	.448	.760	.958
a=1.0	.875	.985	1.00

Note. Each entry is averaged over 2 test lengths and 2 sample size ratios and is based on 400 observations.

As for the main effects, in general, the power of Mantel increased as sample size and item discrimination increased, and its power decreased as sample size ratio changed from 1:1 to 1:2. As with the condition of equal group abilities, there was only little effect of sample size ratio. It was only with the combination of small sample size and low discrimination, and the combination of medium sample size and low discrimination with

sample size ratio of 1:2 that Mantel had a lack of power. Mantel could detect constant uniform DIF with adequate power to perfect accuracy at medium discrimination regardless of sample size.

Unequal Group Ability Distributions **(R-*N*(0,1), F-*N*(-1,1))**

The Type I error and power (α=.05) of the Mantel procedure in detecting uniform DIF (constant) for the condition of large group ability difference are reported in Table 10.

Table 10: The Type I Error and Power of the Mantel Procedure in Detecting Uniform DIF (Constant) for the Condition of Unequal Group Ability Distributions (R-*N*(0,1), F-*N*(-1,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.01	**0.12**	**0.37**		0.39	**0.94**	**1.00**
	600/600		0.04	**0.32**	**0.72**		0.72	**1.00**	**1.00**
	1200/1200		0.04	**0.48**	**0.98**		0.97	**1.00**	**1.00**
1:2 Ratio	200/400		0.03	**0.16**	**0.37**		0.45	**0.86**	**1.00**
	400/800		0.05	**0.36**	**0.65**		0.72	**1.00**	**1.00**
	800/1600		0.08	**0.53**	**0.94**		0.94	**1.00**	**1.00**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.02	**0.06**	**0.17**		0.53	**0.88**	**1.00**
	600/600		0.04	**0.06**	**0.35**		0.69	**0.99**	**1.00**
	1200/1200		0.04	**0.25**	**0.58**		0.98	**1.00**	**1.00**
1:2 Ratio	200/400		0.05	**0.07**	**0.10**		0.43	**0.87**	**0.99**
	400/800		0.05	**0.13**	**0.34**		0.75	**0.99**	**1.00**
	800/1600		0.05	**0.18**	**0.49**		0.96	**1.00**	**1.00**

When the focal group ability was substantially lower than that of the reference group, the Type I error rates of Mantel increased substantially. Its Type I error was in good control when a=0.5. However, most of its Type I error rates were highly inflated (>.10) with a=1.0 or 1.5 (with rates as high as .98), although the inflation was less

excessive with the longer test. Obviously, Mantel detected group ability difference (impact) rather than DIF. Its power under these conditions was not considered. The outcomes at a=1.0 and a=1.5 in Table 10 were highlighted and were not included in the logit analysis because it would be meaningless to further assess the effects of other factors given the unacceptably high Type I error rates for these two levels of item discrimination although there were three exceptions, each with excellent power. Therefore, the logit analysis was conducted for the outcome only for a=0.5 and the independent variables included only the test length (L), sample size (S), and sample size ratio (R).

The result of the logit analysis indicated that for Type I error, a model with only the dependent variable was sufficient to explain the distribution of frequencies observed for the outcome for Mantel (χ^2=9.04, df=11, p=.621). Neither test length nor sample size nor sample size ratio had any effect on the Type I error of Mantel.

For power, a model with the main effect of sample size was sufficient to explain the frequencies of the outcome variable (χ^2= 7.49, df=9, p=.586). Higher power was associated with larger sample size. The Type I error of Mantel was acceptable only when item discrimination was low; however, its power at low discrimination was inadequate at small sample size.

Uniform DIF (Balanced)

The rejection rates (α=.05) of the Mantel procedure for the uniform-DIF (balanced) condition for each combination of ability distribution difference, test length,

sample size, sample size ratio, and item discrimination are presented in Table 11. The Type I error rates are also presented for the ease of reference. From Table 11, one can see that Mantel was not powerful in detecting uniform DIF of balanced type. Its detection rates were very similar to its Type I error rates. When two groups had equal abilities, Mantel failed to detect items with uniform DIF (balanced) almost completely (with detection rates ranging from .03 to .11). However, its detection rates increased as the between-group ability difference increased. When there was a substantial group ability difference, its detection rates could even reach 100% for large sample size and high discrimination. No logit analysis was conducted since the detection rates of Mantel for balanced uniform DIF were basically reflections of its Type I error pattern.

Nonuniform DIF

The rejection rates (α=.05) of the Mantel procedure for the nonuniform-DIF condition for each combination of ability distribution difference, test length, sample size, sample size ratio, and *a*-parameter difference are presented in Table 12 along with its Type I error rates. As with the uniform-DIF (balanced) condition, the detection rates of Mantel for the nonuniform-DIF condition were very similar to its Type I error rates. When reference and focal groups had equal abilities, Mantel failed to detect items with nonuniform DIF completely (with detection rates ranging from .01 to .11). Mantel was not useful for detecting nonuniform DIF. This was expected since Mantel is not designed to detect nonuniform DIF. However, the detection rates increased as the between-group

Table 11

The Rejection Rates of the Mantel Procedure for the Uniform-DIF (Balanced) Condition

		Type I Error			Power		
R-N(0,1), F-N(0,1)							
20-item Test		Item Discrimination			Item Discrimination		
Sample Size		a=0.5	a=1.0	a=1.5	a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300	0.03	0.05	0.03	0.05	0.06	0.05
	600/600	0.05	0.06	0.06	0.06	0.05	0.06
	1200/1200	0.07	0.05	0.04	0.05	0.05	0.08
1:2 Ratio	200/400	0.03	0.04	0.06	0.07	0.04	0.11
	400/800	0.05	0.06	0.04	0.00	0.06	0.05
	800/1600	0.06	0.04	0.07	0.03	0.07	0.06
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.05	0.07	0.04	0.05	0.05	0.03
	600/600	0.04	0.04	0.05	0.07	0.07	0.04
	1200/1200	0.03	0.07	0.03	0.03	0.07	0.03
1:2 Ratio	200/400	0.06	0.07	0.06	0.06	0.03	0.08
	400/800	0.03	0.04	0.06	0.07	0.07	0.05
	800/1600	0.06	0.07	0.04	0.05	0.03	0.03
R-N(0,1), F-N(-.5,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.02	0.04	0.19	0.10	0.10	0.26
	600/600	0.02	0.06	0.20	0.05	0.20	0.41
	1200/1200	0.08	0.18	0.49	0.07	0.23	0.66
1:2 Ratio	200/400	0.04	0.09	0.15	0.04	0.08	0.28
	400/800	0.04	0.07	0.27	0.06	0.22	0.45
	800/1600	0.06	0.14	0.49	0.05	0.32	0.64
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.03	0.06	0.13	0.10	0.09	0.16
	600/600	0.05	0.08	0.09	0.03	0.20	0.26
	1200/1200	0.09	0.08	0.17	0.06	0.23	0.45
1:2 Ratio	200/400	0.03	0.05	0.08	0.07	0.14	0.18
	400/800	0.04	0.08	0.14	0.04	0.17	0.27
	800/1600	0.05	0.08	0.11	0.06	0.31	0.57
R-N(0,1), F-N(-1,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.01	0.12	0.37	0.09	0.28	0.76
	600/600	0.04	0.32	0.72	0.12	0.50	0.95
	1200/1200	0.04	0.48	0.98	0.13	0.85	1.00
1:2 Ratio	200/400	0.03	0.16	0.37	0.07	0.42	0.70
	400/800	0.05	0.36	0.65	0.09	0.70	0.96
	800/1600	0.08	0.53	0.94	0.13	0.93	1.00
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.02	0.06	0.17	0.03	0.16	0.51
	600/600	0.04	0.06	0.35	0.04	0.36	0.72
	1200/1200	0.04	0.25	0.58	0.05	0.61	0.97
1:2 Ratio	200/400	0.05	0.07	0.10	0.10	0.22	0.43
	400/800	0.05	0.13	0.34	0.12	0.37	0.80
	800/1600	0.05	0.18	0.49	0.14	0.56	0.96

Table 12

The Rejection Rates of the Mantel Procedure for the Nonuniform-DIF Condition

		Type I Error			Power		
R-N(0,1), F-N(0,1)							
20-item Test		Item Discrimination			a-Parameter Difference		
Sample Size		a=0.5	a=1.0	a=1.5	a_D=0.5	a_D=1.0	a_D=1.5
1:1 Ratio	300/300	0.03	0.05	0.03	0.06	0.02	0.05
	600/600	0.05	0.06	0.06	0.04	0.01	0.05
	1200/1200	0.07	0.05	0.04	0.04	0.03	0.04
1:2 Ratio	200/400	0.03	0.04	0.06	0.07	0.02	0.08
	400/800	0.05	0.06	0.04	0.08	0.08	0.04
	800/1600	0.06	0.04	0.07	0.07	0.06	0.09
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.05	0.07	0.04	0.09	0.05	0.04
	600/600	0.04	0.04	0.05	0.02	0.08	0.05
	1200/1200	0.03	0.07	0.03	0.07	0.03	0.07
1:2 Ratio	200/400	0.06	0.07	0.06	0.06	0.07	0.08
	400/800	0.03	0.04	0.06	0.06	0.10	0.11
	800/1600	0.06	0.07	0.04	0.13	0.11	0.16
R-N(0,1), F-N(-.5,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.02	0.04	0.19	0.04	0.05	0.10
	600/600	0.02	0.06	0.20	0.04	0.08	0.08
	1200/1200	0.08	0.18	0.49	0.10	0.16	0.12
1:2 Ratio	200/400	0.04	0.09	0.15	0.12	0.18	0.07
	400/800	0.04	0.07	0.27	0.15	0.17	0.19
	800/1600	0.06	0.14	0.49	0.09	0.27	0.17
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.03	0.06	0.13	0.09	0.14	0.13
	600/600	0.05	0.08	0.09	0.10	0.12	0.28
	1200/1200	0.09	0.08	0.17	0.15	0.35	0.40
1:2 Ratio	200/400	0.03	0.05	0.08	0.14	0.15	0.19
	400/800	0.04	0.08	0.14	0.15	0.25	0.36
	800/1600	0.05	0.08	0.11	0.32	0.50	0.52
R-N(0,1), F-N(-1,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.01	0.12	0.37	0.09	0.14	0.14
	600/600	0.04	0.32	0.72	0.05	0.11	0.26
	1200/1200	0.04	0.48	0.98	0.19	0.27	0.33
1:2 Ratio	200/400	0.03	0.16	0.37	0.08	0.17	0.25
	400/800	0.05	0.36	0.65	0.12	0.23	0.32
	800/1600	0.08	0.53	0.94	0.22	0.48	0.50
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.02	0.06	0.17	0.12	0.15	0.30
	600/600	0.04	0.06	0.35	0.25	0.42	0.49
	1200/1200	0.04	0.25	0.58	0.44	0.61	0.79
1:2 Ratio	200/400	0.05	0.07	0.10	0.21	0.30	0.46
	400/800	0.05	0.13	0.34	0.24	0.52	0.70
	800/1600	0.05	0.18	0.49	0.46	0.80	0.91

ability difference increased, as did the Type I error rates. No logit analysis was conducted because the detection rates obviously reflected Mantel's pattern of Type I error rates.

In summary, Mantel had very good control over Type I error when groups were equal in ability or when groups differed in ability but item discrimination was low; its Type I error was inflated but still acceptable when there was a small group ability difference and item discrimination was medium (except for large sample size and shorter test length); under the above conditions, Mantel could be used in detecting constant uniform DIF with adequate to extremely high power even with small sample size (except when item discrimination was low). Mantel is not appropriate to use with the combination of small group ability difference and high discrimination and the combination of large group ability difference and medium to high discrimination because its Type I error was generally unacceptable under these conditions. Mantel was not useful in detecting balanced uniform DIF or nonuniform DIF; its detection rates for these two types of DIF were very similar to its Type I error rates. It might produce spurious results when group ability difference existed.

The GMH Procedure

Uniform DIF (Constant)

Equal Group Ability Distributions **(R-*N*(0,1), F-*N*(0,1))**

The result of logit analysis indicated that for Type I error, a model with only the dependent variable (i.e., the outcome of reject or do not reject at α=.05) was sufficient to explain the distribution of frequencies observed for GMH (χ^2=24.55, df=35, p=.906).

Therefore, none of the independent variables (i.e., L, A, S, and R) had any effect on the Type I error rates of GMH.

For power, the result of logit analysis indicated that a model with A, S, R, and AxS effects was needed to explain the outcome for GMH (χ^2=27.7, df=26, p=.373). The detection rates for the AxS interaction for GMH under the condition of equal group abilities are presented in Table 13. As can be seen from the table, the effect of item discrimination depends on the size of the sample and the difference among the three levels of item discrimination decreased as sample size increased. Much of this ordinal interaction can be explained by the ceiling effect observed at higher discrimination and larger sample size.

Table 13: Detection Rates for the Item Discrimination by Sample Size Interaction for the Uniform-DIF (Constant) Condition for the GMH Procedure (R-N(0,1), F-N(0,1))

	Sample Size		
Item Discrimination	300/300 200/400	600/600 400/800	1200/1200 800/1600
a=0.5	.368	.625	.923
a=1.0	.603	.955	1.00
a=1.5	.850	.993	1.00

Note. Each entry is averaged over 2 test lengths and 2 sample size ratios and is based on 400 observations.

As for the main effects of A, S, and R, the power of GMH increased as sample size and item discrimination increased, its power decreased slightly as sample size ratio changed from 1:1 to 1:2. Although included in the logit model, the effect of sample size ratio was trivial.

The Type I error rates and power (α=.05) of the GMH procedure in detecting uniform DIF (constant) for the condition of equal group abilities are reported in Table 14.

Table 14: The Type I Error and Power of the GMH Procedure in Detecting Uniform DIF (Constant) for the Condition of Equal Group Ability Distributions (R-N(0,1), F-N(0,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.06	0.03	0.02		0.38	0.62	0.85
	600/600		0.06	0.05	0.07		0.62	0.98	0.98
	1200/1200		0.06	0.02	0.05		0.91	1.00	1.00
1:2 Ratio	200/400		0.05	0.06	0.05		0.38	0.59	0.83
	400/800		0.04	0.06	0.04		0.58	0.91	1.00
	800/1600		0.07	0.06	0.06		0.97	1.00	1.00
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.04	0.02	0.04		0.39	0.64	0.87
	600/600		0.02	0.06	0.04		0.70	0.98	1.00
	1200/1200		0.07	0.07	0.06		0.97	1.00	1.00
1:2 Ratio	200/400		0.04	0.06	0.03		0.32	0.56	0.85
	400/800		0.04	0.06	0.01		0.60	0.95	0.99
	800/1600		0.03	0.03	0.04		0.84	1.00	1.00

As can be seen from Table 14, Type I error rates of GMH were within or close to the expected theoretical value of .05. The power of GMH can be best assessed when there was no group ability difference because of its good Type I error control under this condition. As has been explained earlier, statistical power above .70 was considered to be adequate and power over .80 was considered to be high in this study. Based on these criteria, GMH was very powerful with the combination of medium to large sample size and medium to high discrimination, the combination of small sample size and high discrimination, and the combination of large sample size and low discrimination. Under

these conditions, the detection rates ranged from .83 to 1.0. However, the power of GMH was inadequate with the combination of low discrimination and small to medium sample size. It also appeared that sample sizes of 300/300 and 200/400 were not enough for GMH to have adequate power unless item discrimination was high; the power at low discrimination was unsatisfactory except with large sample size.

***Unequal Group Ability Distributions* (R-*N*(0,1), F-*N*(-.5,1))**

The Type I error rates and power (α=.05) of the GMH procedure in detecting uniform DIF (constant) for the condition of small group ability difference are reported in Table 15.

Table 15: The Type I Error and Power of the GMH Procedure in Detecting Uniform DIF (Constant) for the Condition of Unequal Group Ability Distributions (R-*N*(0,1), F-*N*(-.5,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.02	0.04	**0.11**		0.34	0.82	**0.91**
	600/600		0.03	0.02	**0.13**		0.71	0.95	**1.00**
	1200/1200		0.06	0.14	**0.30**		0.89	1.00	**1.00**
1:2 Ratio	200/400		0.05	0.05	**0.08**		0.17	0.74	**0.95**
	400/800		0.08	0.09	**0.17**		0.54	0.95	**1.00**
	800/1600		0.04	0.08	**0.32**		0.88	0.99	**1.00**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.01	0.05	**0.09**		0.42	0.63	**0.95**
	600/600		0.10	0.08	**0.03**		0.63	0.99	**1.00**
	1200/1200		0.07	0.08	**0.12**		0.95	1.00	**1.00**
1:2 Ratio	200/400		0.05	0.03	**0.05**		0.31	0.71	**0.89**
	400/800		0.03	0.07	**0.11**		0.53	0.98	**1.00**
	800/1600		0.02	0.07	**0.11**		0.88	1.00	**1.00**

In general, Type I error was acceptable (<.10) when item discrimination was low and medium with one exception; it also might be acceptable when item discrimination was high but test length was long; but it was unacceptable (>.10) when item discrimination was high and test length was short. Only two levels of item discrimination (a=0.5 and a=1.0) were used in the logit analysis for the condition of small group ability difference. The column under a=1.5 was highlighted and was not used in the logit analysis. From Table 15, one can see that under the condition of small group ability difference (excluding the column with a=1.5), GMH had adequate to extremely high power in detecting constant uniform DIF when both sample size and item discrimination were medium to large; when sample size was large and item discrimination was low, or when sample size was small but item discrimination was medium to high (with one exception). Sample sizes of 300/300 and 200/400 appeared to be inadequate especially when item discrimination was low to medium.

The result of the logit analysis indicated that for Type I error, a model with the main effects of A and S was adequate to explain the frequency distribution of the dependent variable for GMH (χ^2=27.26, df=20, p=.148). Increasing sample size and item discrimination led to an increase in Type I error rates.

For power, a model including the three-way interaction LxSxA as well as the two-way interactions LxS, LxA, AxS, AxR and main effects of A, L, S, and R had to be used to sufficiently explain the outcome variable for GMH (χ^2=8.42, df=10, p=.588). Among the interactions, only AxS had standardized parameter estimates greater than $|1.96|$; therefore, only this interaction is presented in Table 16. As can be seen from the Table

16, the effect of item discrimination depends on the size of the sample and the difference between the two levels of item discrimination decreased as sample size increased. This was due to the ceiling effect observed at large sample size.

Table 16: Detection Rates for the Item Discrimination by Sample Size Interaction for the Uniform-DIF (Constant) Condition for the GMH Procedure (R-N(0,1), F-N(-.5,1))

Item Discrimination	Sample Size		
	300/300 200/400	600/600 400/800	1200/1200 800/1600
a=0.5	.310	.603	.900
a=1.0	.725	.968	.998

Note. Each entry is averaged over 2 test lengths and 2 sample size ratios and is based on 400 observations.

Unequal Group Ability Distributions **(R-N(0,1), F-N(-1,1))**

The Type I error rates and power (α=.05) of the GMH procedure in detecting uniform DIF (constant) for the condition of large group ability difference are reported in Table 17.

When the focal group ability was substantially lower than that of the reference group, the Type I error rates of GMH increased substantially. Its Type I error was acceptable (<.10) when a=0.5. However, its Type I error was generally highly inflated (>.10) when a=1.0 or 1.5 (with three exceptions) although the inflation was less excessive with the longer test. The power under these conditions was not credible, and therefore the outcomes at a=1.0 and a=1.5 were not included in the logit analysis. This means that the

independent variables for the logit analysis only included test length, sample size, and sample size ratio, but not item discrimination.

Table 17: The Type I Error and Power of the GMH Procedure in Detecting Uniform DIF (Constant) for the Condition of Unequal Group Ability Distributions (R-$N(0,1)$, F-$N(-1,1)$)

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			α=0.5	α=1.0	α=1.5		α=0.5	α=1.0	α=1.5
1:1 Ratio	300/300		0.03	**0.14**	**0.30**		0.26	**0.91**	**1.00**
	600/600		0.02	**0.24**	**0.60**		0.58	**1.00**	**1.00**
	1200/1200		0.05	**0.38**	**0.94**		0.88	**1.00**	**1.00**
1:2 Ratio	200/400		0.05	**0.11**	**0.26**		0.25	**0.73**	**0.98**
	400/800		0.04	**0.22**	**0.51**		0.59	**0.98**	**1.00**
	800/1600		0.05	**0.47**	**0.91**		0.85	**1.00**	**1.00**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.06	**0.09**	**0.09**		0.31	**0.75**	**0.96**
	600/600		0.07	**0.08**	**0.29**		0.55	**0.98**	**1.00**
	1200/1200		0.06	**0.22**	**0.41**		0.96	**1.00**	**1.00**
1:2 Ratio	200/400		0.07	**0.06**	**0.06**		0.21	**0.67**	**0.99**
	400/800		0.04	**0.13**	**0.21**		0.59	**0.93**	**1.00**
	800/1600		0.08	**0.19**	**0.36**		0.84	**1.00**	**1.00**

The result of the logit analysis indicated that for Type I error, a model with only the dependent variable was sufficient to explain the distribution of frequencies observed for the outcome for GMH (χ^2=7.31, df=11, p=.774). Neither test length nor sample size nor sample size ratio had any effect on the Type I error of GMH.

For power, a model with only the main effect of sample size was sufficient (χ^2=13.17, df=9, p=.155) and the power of GMH increased with the increase of sample size. The Type I error of GMH was only acceptable when item discrimination was low;

however, its power at low discrimination was inadequate except when sample size was large.

Uniform DIF (Balanced)

Equal Group Ability Distributions (R-$N(0,1)$, F-$N(0,1)$)

The result of the logit analysis indicated that a model with the effects of A, S, and AxS was adequate to explain the results for GMH (χ^2=31.46, df=27, p=.253). In Table 18 the detection rates for the AxS interaction for uniform-DIF (balanced) condition for GMH when two groups had equal abilities are presented. As can be seen from the table, the difference between a=0.5 and a=1 increased as sample size increased, but the difference between a=1 and a=1.5 decreased as sample size increased. This ordinal interaction basically reflects a ceiling effect at larger sample size and higher discrimination. As for the main effects, the power increased as item discrimination and sample size increased.

Table 18: Detection Rates for the Item Discrimination by Sample Size Interaction for the Uniform-DIF (Balanced) Conditions for the GMH Procedure (R-$N(0,1)$, F-$N(0,1)$)

Item Discrimination	Sample Size		
	300/300 200/400	600/600 400/800	1200/1200 800/1600
a=0.5	.160	.223	.488
a=1.0	.320	.618	.888
a=1.5	.595	.895	.993

Note. Each entry is averaged over 2 test lengths and 2 sample size ratios and is based on 400 observations.

The rejection rates (α=.05) of the GMH procedure for the uniform-DIF (balanced) condition for the condition of equal group abilities are presented in Table 19 along with its Type I error. Since the Type I error of GMH was in good control when there was no group ability difference, its power in detecting balanced uniform DIF can be best assessed for this condition.

Table 19: The Type I Error and Power of the GMH Procedure in Detecting Uniform DIF (Balanced) for the Condition of Equal Group Ability Distributions (R-N(0,1), F-N(0,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.06	0.03	0.02		0.12	0.37	0.57
	600/600		0.06	0.05	0.07		0.29	0.67	0.93
	1200/1200		0.06	0.02	0.05		0.49	0.88	1.00
1:2 Ratio	200/400		0.05	0.06	0.05		0.13	0.28	0.59
	400/800		0.04	0.06	0.04		0.18	0.57	0.86
	800/1600		0.07	0.06	0.06		0.48	0.86	0.98
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.04	0.02	0.04		0.21	0.32	0.68
	600/600		0.02	0.06	0.04		0.21	0.64	0.94
	1200/1200		0.07	0.07	0.06		0.49	0.93	1.00
1:2 Ratio	200/400		0.04	0.06	0.03		0.18	0.31	0.54
	400/800		0.04	0.06	0.01		0.21	0.59	0.85
	800/1600		0.03	0.03	0.04		0.49	0.88	0.99

As can be seen from Table 19, GMH was very powerful in detecting balanced uniform DIF with the combination of large sample size and high discrimination or combination of high discrimination and medium sample size (with detection rates ranging from .85 to 1.0), but its power was inadequate, i.e., <.70, in all the other conditions and

even very low when item discrimination was low or when sample size was small.

Unequal Group Ability Distributions **(R-*N*(0,1), F-*N*(-.5,1))**

Only two levels of item discrimination (a=0.5 and 1.0) were included in the logit analysis for the condition of small group ability difference as most of the Type I error rates for a=1.5 exceeded the acceptable rate of <.10. The result of logit analysis indicated that a model including A, S, R, and AxS effects adequately explained the frequency distribution of the dependent variable for GMH (χ^2=20.58, df=17, p=.246). The detection rates for the AxS interaction for uniform-DIF (balanced) condition for GMH are presented in Table 20. As can be seen from the table, the difference between a=0.5 and a=1 increased as sample size increased.

Table 20: Detection Rates for the Item Discrimination by Sample Size Interaction for the Uniform-DIF (Balanced) Condition for the GMH Procedure (R-*N*(0,1), F-*N*(-.5,1))

	Sample Size		
Item Discrimination	300/300 200/400	600/600 400/800	1200/1200 800/1600
a=0.5	.170	.400	.695
a=1.0	.270	.703	.960

Note. Each entry is averaged over 2 test lengths and 2 sample size ratios and is based on 400 observations.

The rejection rates (α=.05) of the GMH procedure for the uniform-DIF (balanced) condition for the condition of small group ability difference are presented in Table 21 along with its Type I error. The column not included in the logit analysis was highlighted. Increasing sample size and item discrimination generally led to an increase in detection rates, and the sample size ratio of 1:2 generally produced lower detection rates than the

ratio of 1:1. The power of GMH was adequate with the combination of medium item discrimination and medium sample size and was adequate with medium item discrimination and medium sample size for short test length. It was low with low item discrimination or small sample size.

Table 21: The Type I Error and Power of the GMH Procedure in Detecting Uniform DIF (Balanced) for the Condition of Unequal Group Ability Distributions (R-N(0,1), F-N(-.5,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.02	0.04	**0.11**		0.20	0.37	**0.74**
	600/600		0.03	0.02	**0.13**		0.38	0.73	**0.96**
	1200/1200		0.06	0.14	**0.30**		0.63	0.95	**1.00**
1:2 Ratio	200/400		0.05	0.05	**0.08**		0.19	0.42	**0.73**
	400/800		0.08	0.09	**0.17**		0.21	0.74	**0.99**
	800/1600		0.04	0.08	**0.32**		0.53	0.97	**1.00**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.01	0.05	**0.09**		0.20	0.44	**0.73**
	600/600		0.10	0.08	**0.03**		0.29	0.71	**0.96**
	1200/1200		0.07	0.08	**0.12**		0.53	0.98	**1.00**
1:2 Ratio	200/400		0.05	0.03	**0.05**		0.09	0.37	**0.58**
	400/800		0.03	0.07	**0.11**		0.20	0.63	**0.93**
	800/1600		0.02	0.07	**0.11**		0.47	0.94	**1.00**

Unequal Group Ability Distributions **(R-N(0,1), F-N(-1,1))**

For the condition of large group ability difference, only the outcome at a=0.5 was included in the logit analysis. As most of the Type I error rates at a=1.0 and 1.5 exceeded the acceptable rate of <.10, this caused one to question the power other than when a=0.5. The independent variables for the logit analysis only included test length, sample size,

and sample size ratio. The result of logit analysis indicated that a model with the main effect of sample size was sufficient to explain the outcome (χ^2=11.96, df=9, p=.216). The rejection rates (α=.05) of the GMH procedure for the uniform-DIF (balanced) condition for the condition of large group ability difference are presented in Table 22 along with its Type I error. The columns not included in the logit analysis were highlighted. As shown in Table 22, the power of GMH was very low when item discrimination was low although the power increased as sample size increased. The power of GMH was very high when item discrimination was medium to high, so did its Type I error.

Table 22: The Type I Error and Power of the GMH Procedure in Detecting Uniform DIF (Balanced) for the Condition of Unequal Group Ability Distributions (R-N(0,1), F-N(-1,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.03	**0.14**	**0.30**		0.12	**0.56**	**0.94**
	600/600		0.02	**0.24**	**0.60**		0.32	**0.85**	**1.00**
	1200/1200		0.05	**0.38**	**0.94**		0.65	**1.00**	**1.00**
1:2 Ratio	200/400		0.05	**0.11**	**0.26**		0.20	**0.54**	**0.85**
	400/800		0.04	**0.22**	**0.51**		0.24	**0.89**	**1.00**
	800/1600		0.05	**0.47**	**0.91**		0.57	**1.00**	**1.00**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.06	**0.09**	**0.09**		0.13	**0.43**	**0.80**
	600/600		0.07	**0.08**	**0.29**		0.28	**0.78**	**0.97**
	1200/1200		0.06	**0.22**	**0.41**		0.55	**0.95**	**1.00**
1:2 Ratio	200/400		0.07	**0.06**	**0.06**		0.20	**0.44**	**0.67**
	400/800		0.04	**0.13**	**0.21**		0.26	**0.69**	**0.95**
	800/1600		0.08	**0.19**	**0.36**		0.48	**0.98**	**1.00**

Nonuniform DIF

Equal Group Ability Distributions (R-*N*(0,1), F-*N*(0,1))

The result of logit analysis indicated that a model including the interaction (A_DxS) as well as the main effects (A_D, L, S) was needed to explain the variations in the dependent variable (χ^2=24.98, df=26, p=.52). The detection rates for the A_DxS interaction are presented in Table 23. As can be seen from the table, the difference between a_D=0.5 and a_D=1.0 increased as sample size increased but the difference between a_D=1.0 and a_D=1.5 was relatively uniform across sample sizes.

Table 23: Detection Rates for the *a*-Parameter Difference by Sample Size Interaction for the Nonuniform-DIF Condition for the GMH Procedure (R-*N*(0,1), F-*N*(0,1))

	Sample Size		
a-Parameter Difference	300/300 200/400	600/600 400/800	1200/1200 800/1600
a_D =0.5	.085	.160	.315
a_D =1.0	.220	.415	.748
a_D =1.5	.345	.613	.915

Note. Each entry is averaged over 2 test lengths and 2 ratios and is based on 400 observations.

The rejection rates (α=.05) of the GMH procedure for the nonuniform DIF condition for the condition of equal group abilities are presented in Table 24 along with its Type I error rates. Except when sample size was large and item discrimination was medium to high, the power of GMH in detecting nonuniform DIF was very low in all the other conditions. Power generally increased with the increase of *a*-parameter difference, test length, and sample size.

Table 24: The Type I Error and Power of the GMH Procedure in Detecting Nonuniform DIF for the Condition of Equal Group Ability Distributions (R-$N(0,1)$, F-$N(0,1)$)

			Type I Error				Power		
20-item Test			Item Discrimination				a-Parameter Difference		
Sample Size			a=0.5	a=1.0	a=1.5		a_D=0.5	a_D=1.0	a_D=1.5
1:1 Ratio	300/300		0.06	0.03	0.02		0.07	0.17	0.29
	600/600		0.06	0.05	0.07		0.16	0.37	0.58
	1200/1200		0.06	0.02	0.05		0.29	0.63	0.89
1:2 Ratio	200/400		0.05	0.06	0.05		0.11	0.21	0.35
	400/800		0.04	0.06	0.04		0.19	0.41	0.52
	800/1600		0.07	0.06	0.06		0.27	0.70	0.87
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.04	0.02	0.04		0.09	0.26	0.36
	600/600		0.02	0.06	0.04		0.14	0.44	0.70
	1200/1200		0.07	0.07	0.06		0.33	0.79	0.96
1:2 Ratio	200/400		0.04	0.06	0.03		0.07	0.24	0.38
	400/800		0.04	0.06	0.01		0.15	0.44	0.65
	800/1600		0.03	0.03	0.04		0.37	0.87	0.94

Unequal Group Ability Distributions **(R-$N(0,1)$, F-$N(-.5,1)$)**

For the condition of small group ability difference, only two levels of item discrimination (a=0.5 and 1.0) were included in the logit analysis. As most of the Type I error rates at a=1.5 were not acceptable, the power under this condition was questionable; therefore, a=1.5 was not included in the logit analysis.

The result of logit analysis indicated that a model including the interaction (RxS) as well as the main effects (R, L, S) was needed to explain the variations in the dependent variable (χ^2=17.7, df=17, p=.408). The standardized parameter estimates for the RxS interaction were smaller than $|1.96|$; therefore, no further explanations are provided here.

The rejection rates (α=.05) of the GMH procedure for the nonuniform DIF condition for the condition of small group ability difference are presented in Table 25

along with its Type I error rates. The power of GMH was very low in most conditions except when sample size was large and item discrimination was medium; however, its power generally increased as sample size, sample size ratio, and test length increased.

Table 25: The Type I Error and Power of the GMH Procedure in Detecting Nonuniform DIF for the Condition of Unequal Group Ability Distributions (R-N(0,1), F-N(-.5,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				a-Parameter Difference		
Sample Size			a=0.5	a=1.0	a=1.5		a_D=0.5	a_D=1.0	a_D=1.5
1:1 Ratio	300/300		0.02	0.04	**0.11**		0.13	0.31	**0.45**
	600/600		0.03	0.02	**0.13**		0.17	0.52	**0.67**
	1200/1200		0.06	0.14	**0.30**		0.40	0.78	**0.97**
1:2 Ratio	200/400		0.05	0.05	**0.08**		0.15	0.30	**0.45**
	400/800		0.08	0.09	**0.17**		0.30	0.54	**0.73**
	800/1600		0.04	0.08	**0.32**		0.34	0.84	**0.99**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.01	0.05	**0.09**		0.17	0.25	**0.41**
	600/600		0.10	0.08	**0.03**		0.32	0.54	**0.77**
	1200/1200		0.07	0.08	**0.12**		0.48	0.89	**0.99**
1:2 Ratio	200/400		0.05	0.03	**0.05**		0.22	0.29	**0.46**
	400/800		0.03	0.07	**0.11**		0.27	0.59	**0.80**
	800/1600		0.02	0.07	**0.11**		0.51	0.90	**0.96**

Unequal Group Ability Distributions **(R-N(0,1), F-N(-1,1))**

For the condition of large group ability difference, only the outcome at a=0.5 was included in the logit analysis. As the Type I error rates at a=1.0 and 1.5 were far above the acceptable rate, >.10, the power under these conditions was considered, and therefore was not included in the logit analysis. The results of logit analysis indicated that a model including the main effect of sample size was sufficient (χ^2=6.84, df=9, p=.654). Increasing sample size led to an increase in power.

The rejection rates (α=.05) of the GMH procedure for the nonuniform DIF condition for the condition of large group ability difference are presented in Table 26 along with its Type I error rates. The Type I error of GMH was only acceptable when item discrimination was low, however, its power under this condition was inadequate.

Table 26: The Type I Error and Power of the GMH Procedure in Detecting Nonuniform DIF for the Condition of Unequal Group Ability Distributions (R-N(0,1), F-N(-1,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				a-Parameter Difference		
Sample Size			a=0.5	a=1.0	a=1.5		a_D=0.5	a_D=1.0	a_D=1.5
1:1 Ratio	300/300		0.03	**0.14**	**0.30**		0.18	**0.44**	**0.55**
	600/600		0.02	**0.24**	**0.60**		0.37	**0.70**	**0.89**
	1200/1200		0.05	**0.38**	**0.94**		0.63	**0.91**	**0.99**
1:2 Ratio	200/400		0.05	**0.11**	**0.26**		0.14	**0.41**	**0.68**
	400/800		0.04	**0.22**	**0.51**		0.32	**0.70**	**0.86**
	800/1600		0.05	**0.47**	**0.91**		0.55	**0.88**	**0.98**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.06	**0.09**	**0.09**		0.15	**0.42**	**0.56**
	600/600		0.07	**0.08**	**0.29**		0.37	**0.64**	**0.87**
	1200/1200		0.06	**0.22**	**0.41**		0.58	**0.96**	**1.00**
1:2 Ratio	200/400		0.07	**0.06**	**0.06**		0.21	**0.46**	**0.64**
	400/800		0.04	**0.13**	**0.21**		0.39	**0.72**	**0.91**
	800/1600		0.08	**0.19**	**0.36**		0.67	**0.94**	**1.00**

In summary, GMH had very good control over Type I error when groups were equal in ability or when groups differed in ability but item discrimination was low. Its Type I error was inflated but still acceptable when there was a small group ability difference and item discrimination was medium (except for large sample size and shorter test length). Under the above conditions, GMH could be used in detecting constant uniform DIF with adequate to extremely high power when both sample size and item

discrimination were medium to high but it was not powerful with the combination of small to medium sample size and low discrimination. GMH could also be used to detect balanced uniform DIF with high power in some conditions (such as large sample size and medium to high item discrimination), but its power was very low in most other conditions. GMH could also be used to detect nonuniform DIF to some degree; however, its power was low in most conditions except when item discrimination was medium to high and sample size was large. GMH might be used with the condition of small group difference, high discrimination and longer test length. It should not be used with the combination of large group ability difference and medium to high discrimination because its Type I error was generally unacceptable under these conditions.

The STND Procedure

Uniform DIF (Constant)

Equal Group Ability Distributions **(R-*N*(0,1), F-*N*(0,1))**

The result of logit analysis indicated that for Type I error, a model with only the effect of sample size ratio was sufficient to explain the distribution of frequencies observed for STND (χ^2=20.18, df=34, p=.971). Type I error was smaller for sample size ratio of 1:2.

For power, a model including the interaction of AxS as well as main effects A, S, and R was needed to explain the results for STND (χ^2=33.66, df=26, p=.154). The detection rates for the AxS interaction are reported in Table 27. As can be seen from the table, the differences among the three levels of item discrimination decreased as sample

size increased; similarly, the differences across sample sizes decreased as item discrimination increased. This interaction basically reflects the ceiling effect as item discrimination and sample size increased.

Table 27: Detection Rates for the Item Discrimination by Sample Size Interaction for the Uniform-DIF (Constant) Condition for the STND Procedure (R-$N(0,1)$, F-$N(0,1)$)

	Sample Size		
Item Discrimination	300/300 200/400	600/600 400/800	1200/1200 800/1600
a=0.5	.423	.643	.898
a=1.0	.680	.950	.995
a=1.5	.935	1.00	1.00

Note. Each entry is averaged over 2 test lengths and 2 sample size ratios and is based on 400 observations.

The Type I error rates and power of STND in detecting uniform DIF (constant) for the condition of equal group abilities are presented in Table 28. The Type I error rates of STND were within or close to .05 in most conditions; however, it had much lower Type I error rates when the sample size ratio between the focal and reference groups was 1:2 (mean error rate of .012) than when the ratio was 1:1 (mean error rate of .054). The power of STND generally increased as sample size and item discrimination increased and decreased as sample size ratio changed from 1:1 to 1:2. STND was extremely powerful in detecting constant uniform DIF with the combination of medium to large sample size and medium to high discrimination, the combination of small sample size and high discrimination, and the combination of large sample size and low discrimination. Under these conditions, the power ranged from .85 to 1.0. The power of STND was adequate

Table 28: The Type I Error and Power of the STND Procedure in Detecting Uniform DIF (Constant) for the Condition of Equal Group Ability Distributions (R-$N(0,1)$, F-$N(0,1)$)

			Type I Error				Power		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.05	0.06	0.04		0.56	0.73	0.94
	600/600		0.06	0.05	0.06		0.75	0.97	0.99
	1200/1200		0.07	0.06	0.03		0.96	1.00	1.00
1:2 Ratio	200/400		0.01	0.00	0.02		0.25	0.55	0.82
	400/800		0.00	0.01	0.02		0.54	0.88	1.00
	800/1600		0.02	0.01	0.02		0.93	1.00	1.00
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.06	0.07	0.06		0.62	0.82	0.98
	600/600		0.05	0.06	0.05		0.86	1.00	1.00
	1200/1200		0.02	0.07	0.05		1.00	1.00	1.00
1:2 Ratio	200/400		0.02	0.01	0.02		0.26	0.47	0.85
	400/800		0.00	0.02	0.01		0.57	0.95	0.99
	800/1600		0.01	0.02	0.00		0.85	1.00	1.00

with the combination of small sample size, medium discrimination and sample size ratio of 1:1, and the combination of medium sample size, low discrimination and sample size ratio of 1:1. STND had inadequate power (<.70) under other conditions. The effect of sample size ratio on STND was very obvious, it had lower power when the ratio was 1:2 than when it was 1:1. This is consistent with its Type I error rates, which were lower when the ratio was 1:2. Sample sizes of 300/300 and 200/400 appeared to be too small for STND to have adequate power, especially when item discrimination was low.

Unequal Group Ability Distributions **(R-$N(0,1)$, F-$N(-.5,1)$)**

The Type I error rates and power of STND in detecting uniform DIF (constant) for the condition of small group ability difference are presented in Table 29. When there was a small group ability difference, the Type I error rates of STND increased

Table 29: The Type I Error and Power of the STND Procedure in Detecting Uniform DIF (Constant) for the Condition of Unequal Group Ability Distributions (R-$N(0,1)$, F-$N(-.5,1)$)

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.06	0.08	**0.23**		0.54	0.94	**0.99**
	600/600		0.04	0.13	**0.22**		0.85	0.98	**1.00**
	1200/1200		0.07	0.18	**0.49**		0.94	1.00	**1.00**
1:2 Ratio	200/400		0.02	0.01	**0.05**		0.16	0.63	**0.95**
	400/800		0.01	0.03	**0.12**		0.45	0.95	**1.00**
	800/1600		0.01	0.06	**0.27**		0.88	0.99	**1.00**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.09	0.10	**0.15**		0.62	0.91	**0.99**
	600/600		0.06	0.12	**0.16**		0.80	1.00	**1.00**
	1200/1200		0.12	0.10	**0.20**		0.99	1.00	**1.00**
1:2 Ratio	200/400		0.01	0.02	**0.05**		0.27	0.75	**0.91**
	400/800		0.01	0.04	**0.09**		0.46	0.97	**1.00**
	800/1600		0.02	0.02	**0.04**		0.91	1.00	**1.00**

considerably. In general, Type I error was acceptable (<.10) when item discrimination was low to medium except when item discrimination was medium and sample size ratio was 1:1; most of the Type I error was unacceptably high (>.10) when item discrimination was high except with the 40-item test and sample size ratio of 1:2. Consequently, the power for high discrimination was not credible, therefore, the outcome under this condition was not included in the logit analysis. Only two levels of item discrimination (a=0.5 and a=1.0) were used in the logit analysis for the condition of small group ability difference.

The result of the logit analysis indicated that for Type I error, a model with the effects of A, S, and R was needed to explain the outcome (χ^2=13.47, df=19, p=.814). From Table 29, one can see that in addition to the sample size and item discrimination

effects, STND was also affected by sample size ratio. The ratio of 1:2 produced substantially lower Type I error rates than the ratio of 1:1. As a matter of fact, the Type I error rates of STND were within or close to .05 in most conditions when the ratio was 1:2.

For power, the result of logit analysis indicated that a model including the terms of A, S, R, and AxS was needed to explain the results for STND (χ^2=20.79, df=17, p=.236). The detection rates for the AxS interaction are presented in Table 30. As can be seen from the table, the effect of item discrimination depends on the size of sample; the difference between the two levels of item discrimination decreased as sample size increased. This interaction can be explained by the ceiling effect observed at higher discrimination and larger sample size.

Table 30: Detection Rates for the Item Discrimination by Sample Size Interaction for the Uniform-DIF (Constant) Condition for the STND Procedure (R-N(0,1), F-N(-.5,1))

	Sample Size		
Item Discrimination	300/300 200/400	600/600 400/800	1200/1200 800/1600
a=0.5	.398	.808	.960
a=1.0	.640	.975	1.00

Note. Each entry is averaged over 2 test lengths and 2 sample size ratios and is based on 400 observations.

Like its Type I error, the power of STND increased as sample size and item discrimination increased, and decreased as sample size ratio changed from 1:1 to 1:2. It was only with the combination of small sample size and low discrimination, and the

combination of medium sample size and low discrimination and sample size ratio of 1:2 that STND had a lack of power. STND could detect constant uniform DIF with perfect accuracy at medium to large sample size and medium discrimination (see Table 29). One must be cautious in the power shown for α=1.0 and sample size ratio of 1:1 because of high Type I error.

Unequal Group Ability Distributions **(R-*N*(0,1), F-*N*(-1,1))**

The Type I error rates and power of STND in detecting uniform DIF (constant) for the condition of large group ability difference are presented in Table 31.

Table 31: The Type I Error and Power of the STND Procedure in Detecting Uniform DIF (Constant) for the Condition of Unequal Group Ability Distributions (R-*N*(0,1), F-*N*(-1,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			α=0.5	α=1.0	α=1.5		α=0.5	α=1.0	α=1.5
1:1 Ratio	300/300		0.03	**0.20**	**0.56**		0.51	**0.97**	**1.00**
	600/600		0.12	**0.46**	**0.81**		0.83	**1.00**	**1.00**
	1200/1200		0.09	**0.65**	**1.00**		1.00	**1.00**	**1.00**
1:2 Ratio	200/400		0.02	**0.09**	**0.33**		0.33	**0.76**	**1.00**
	400/800		0.02	**0.19**	**0.53**		0.62	**1.00**	**1.00**
	800/1600		0.05	**0.42**	**0.91**		0.91	**1.00**	**1.00**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.10	**0.18**	**0.27**		0.61	**0.91**	**0.98**
	600/600		0.11	**0.17**	**0.46**		0.70	**0.98**	**1.00**
	1200/1200		0.11	**0.45**	**0.75**		0.98	**1.00**	**1.00**
1:2 Ratio	200/400		0.04	**0.05**	**0.06**		0.32	**0.74**	**0.99**
	400/800		0.02	**0.13**	**0.28**		0.62	**0.97**	**1.00**
	800/1600		0.03	**0.13**	**0.36**		0.87	**1.00**	**1.00**

As can be seen from Table 31, when the focal group ability was substantially lower than that of the reference group, the Type I error rates of STND increased substantially. Its Type I error was below .05 with α=0.5 and sample size ratio of 1:2.

However, its Type I error was highly inflated when a=1.0 or 1.5 (with rates as high as 1.0) with a couple of exceptions. Like Mantel and GMH, STND detected group ability difference (impact) rather than DIF. The outcomes at a=1.0 and a=1.5 were not included in the logit analysis because it would be meaningless to further assess the effects of other factors given the unacceptably high Type I error rates for these two levels of item discrimination. Therefore, the logit analysis was conducted for the outcome only at a=0.5.

The result of the logit analysis indicated that for Type I error, a model with the effect of sample size ratio was adequate (χ^2=10.38, df=10, p=.408); the ratio of 1:2 produced substantially lower Type I error rates than the ratio of 1:1.

For power, a model with the effects of S, R and RxS was needed to explain the outcome (χ^2=10.42, df=6, p=.148). Regarding the RxS interaction, the effect of sample size ratio was more obvious with smaller sample size than with large sample size (see Table 32 below). As the Type I error of STND was actually only acceptable at low discrimination and sample size ratio of 1:2, its power under other conditions was questionable.

Table 32: Detection Rates for the Sample Size Ratio by Sample Size Interaction for the Uniform-DIF (Constant) Condition for the STND Procedure (R-N(0,1), F-N(-1,1))

	Sample Size		
Sample Size Ratio	small	medium	large
1:1	.560	.765	.990
1:2	.325	.620	.890

Note. Each entry is averaged over 2 test lengths and is based on 200 observations.

Uniform DIF (Balanced)

The rejection rates (α=.05) of the STND procedure for the uniform-DIF (balanced) condition along with its Type I error rates for each combination of group ability distribution difference, test length, sample size, and sample size ratio are presented in Table 33.

STND was not useful for detecting uniform DIF of balanced type. When two groups had equal abilities, STND failed to detect items with uniform DIF (balanced) almost completely. However, its detection rates increased as the between-group ability difference increased, as did its Type I error rates. Like the Mantel procedure, it probably detected group ability difference or impact rather than DIF. No logit analysis was conducted for STND for the uniform-DIF (balanced) condition since the results were very similar to its Type I error rates.

Nonuniform DIF

The rejection rates (α=.05) of the STND procedure for the nonuniform DIF condition along with its Type I error rates for each combination of group ability distribution difference, test length, sample size, and sample size ratio are presented in Table 34.

As with the uniform-DIF (balanced) condition, STND was not useful for detecting nonuniform DIF. This was expected since STND is not designed to detect nonuniform DIF. However, the detection rates increased as the between-group ability difference

Table 33

The Rejection Rates of the STND Procedure for the Uniform-DIF (Balanced) Condition

		Type I Error			**Power**		
R-N(0,1), F-N(0,1)							
20-item Test		Item discrimination			Item discrimination		
Sample Size		a=0.5	a=1.0	a=1.5	a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300	0.05	0.06	0.04	0.06	0.08	0.05
	600/600	0.06	0.05	0.06	0.08	0.05	0.06
	1200/1200	0.07	0.06	0.03	0.06	0.04	0.06
1:2 Ratio	200/400	0.01	0.00	0.02	0.03	0.01	0.01
	400/800	0.00	0.01	0.02	0.00	0.01	0.01
	800/1600	0.02	0.01	0.02	0.01	0.00	0.01
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.06	0.07	0.06	0.10	0.06	0.04
	600/600	0.05	0.06	0.05	0.06	0.06	0.05
	1200/1200	0.02	0.07	0.05	0.03	0.07	0.05
1:2 Ratio	200/400	0.02	0.01	0.02	0.02	0.00	0.02
	400/800	0.00	0.02	0.01	0.01	0.01	0.01
	800/1600	0.01	0.02	0.00	0.01	0.01	0.01
R-N(0,1), F-N(-.5,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.06	0.08	0.23	0.09	0.18	0.41
	600/600	0.04	0.13	0.22	0.09	0.28	0.62
	1200/1200	0.07	0.18	0.49	0.12	0.54	0.91
1:2 Ratio	200/400	0.02	0.01	0.05	0.00	0.07	0.17
	400/800	0.01	0.03	0.12	0.01	0.14	0.34
	800/1600	0.01	0.06	0.27	0.00	0.26	0.67
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.09	0.10	0.15	0.16	0.17	0.29
	600/600	0.06	0.12	0.16	0.10	0.30	0.48
	1200/1200	0.12	0.10	0.20	0.12	0.45	0.69
1:2 Ratio	200/400	0.01	0.02	0.05	0.03	0.05	0.12
	400/800	0.01	0.04	0.09	0.03	0.11	0.22
	800/1600	0.02	0.02	0.04	0.01	0.24	0.55
R-N(0,1), F-N(-1,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.03	0.20	0.56	0.20	0.53	0.91
	600/600	0.12	0.46	0.81	0.25	0.76	0.99
	1200/1200	0.09	0.65	1.00	0.30	0.99	1.00
1:2 Ratio	200/400	0.02	0.09	0.33	0.05	0.32	0.63
	400/800	0.02	0.19	0.53	0.08	0.71	0.95
	800/1600	0.05	0.42	0.91	0.12	0.94	1.00
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.10	0.18	0.27	0.13	0.35	0.73
	600/600	0.11	0.17	0.46	0.22	0.65	0.92
	1200/1200	0.11	0.45	0.75	0.31	0.88	1.00
1:2 Ratio	200/400	0.04	0.05	0.06	0.10	0.24	0.48
	400/800	0.02	0.13	0.28	0.09	0.39	0.86
	800/1600	0.03	0.13	0.36	0.14	0.60	0.98

Table 34

The Rejection Rates of the STND Procedure for the Nonuniform Condition

		Type I Error			Power		
R-N(0,1), F-N(0,1)							
20-item Test		Item Discrimination			a-Parameter Difference		
Sample Size		a=0.5	a=1.0	a=1.5	a_D=0.5	a_D=1.0	a_D=1.5
1:1 Ratio	300/300	0.05	0.06	0.04	0.05	0.04	0.04
	600/600	0.06	0.05	0.06	0.05	0.04	0.05
	1200/1200	0.07	0.06	0.03	0.04	0.03	0.06
1:2 Ratio	200/400	0.01	0.00	0.02	0.03	0.00	0.04
	400/800	0.00	0.01	0.02	0.03	0.01	0.03
	800/1600	0.02	0.01	0.02	0.00	0.00	0.01
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.06	0.07	0.06	0.08	0.06	0.05
	600/600	0.05	0.06	0.05	0.08	0.08	0.11
	1200/1200	0.02	0.07	0.05	0.07	0.04	0.09
1:2 Ratio	200/400	0.02	0.01	0.02	0.01	0.04	0.03
	400/800	0.00	0.02	0.01	0.03	0.02	0.03
	800/1600	0.01	0.02	0.00	0.03	0.02	0.04
R-N(0,1), F-N(-.5,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.06	0.08	0.23	0.11	0.10	0.14
	600/600	0.04	0.13	0.22	0.13	0.20	0.25
	1200/1200	0.07	0.18	0.49	0.17	0.37	0.36
1:2 Ratio	200/400	0.02	0.01	0.05	0.07	0.14	0.02
	400/800	0.01	0.03	0.12	0.05	0.16	0.16
	800/1600	0.01	0.06	0.27	0.06	0.22	0.16
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.09	0.10	0.15	0.19	0.28	0.31
	600/600	0.06	0.12	0.16	0.28	0.38	0.56
	1200/1200	0.12	0.10	0.20	0.43	0.71	0.89
1:2 Ratio	200/400	0.01	0.02	0.05	0.08	0.11	0.19
	400/800	0.01	0.04	0.09	0.13	0.30	0.36
	800/1600	0.02	0.02	0.04	0.26	0.54	0.64
R-N(0,1), F-N(-1,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.03	0.20	0.56	0.18	0.31	0.34
	600/600	0.12	0.46	0.81	0.20	0.43	0.67
	1200/1200	0.09	0.65	1.00	0.42	0.65	0.80
1:2 Ratio	200/400	0.02	0.09	0.33	0.06	0.13	0.23
	400/800	0.02	0.19	0.53	0.13	0.28	0.31
	800/1600	0.05	0.42	0.91	0.17	0.51	0.60
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.10	0.18	0.27	0.34	0.55	0.77
	600/600	0.11	0.17	0.46	0.57	0.88	0.96
	1200/1200	0.11	0.45	0.75	0.84	0.98	1.00
1:2 Ratio	200/400	0.04	0.05	0.06	0.24	0.41	0.59
	400/800	0.02	0.13	0.28	0.30	0.71	0.85
	800/1600	0.03	0.13	0.36	0.55	0.92	1.00

increased, so did the Type I error rates. No logit analysis was conducted because the detection rates of STND were reflections of its high Type I error rates.

In summary, STND had very good control over Type I error when groups were equal in ability; its Type I error was inflated but still acceptable when there was a small group ability difference and item discrimination was low to medium or when there was a large group ability difference and item discrimination was low (with a few exceptions). Under the above conditions, STND could be used in detecting constant uniform DIF with adequate to extremely high power (except when item discrimination was low and sample size was small). STND should not be used with the combination of small group ability difference and high discrimination, except when sample size ratio was 1:2 and test length was long and the combination of large group ability difference and medium to high discrimination because its Type I error was generally unacceptably high under these conditions. STND was not useful in detecting balanced uniform DIF or nonuniform DIF and its detection rates for these two types of DIF were very similar to its Type I error rates.

The SIBTEST Procedure

Uniform DIF (Constant)

Equal Group Ability Distributions (R-*N*(0,1), F-*N*(0,1))

The result of logit analysis indicated that for Type I error, a model with only the dependent variable was sufficient to explain the distribution of frequencies observed for

SIBTEST (χ^2=17.02, df=35, p=.996). Therefore, none of the independent variables had any effect on the Type I error rates of SIBTEST.

For power, the result of the logit analysis indicated that a model with the effects of A, L, S, R, AxS, AxR, AxL, LxR, LxS, RxS was needed to explain the outcome for SIBTEST (χ^2=15.58, df=16, p=.483). However, among the six two-way interactions, only AxS and LxR had standardized parameter estimates greater than |1.96| and therefore were not reported. The detection rates for the AxS interaction are presented in Table 35. The differences among the three levels of item discrimination decreased as sample size increased; similarly, the differences across sample sizes decreased as item discrimination increased. Apparently, this interaction reflects the ceiling effect observed at higher discrimination and larger sample size.

Table 35: Detection Rates for the Item Discrimination by Sample Size Interaction for the Uniform-DIF (Constant) Condition for the SIBTEST Procedure ((R-N(0,1), F-N(0,1))

	Sample Size		
Item Discrimination	300/300 200/400	600/600 400/800	1200/1200 800/1600
a=0.5	.535	.765	.958
a=1.0	.758	.973	.998
a=1.5	.948	.995	1.00

Note. Each entry is averaged over 2 test lengths and 2 sample size ratios and is based on 400 observations

The detection rates for the LxR interaction are presented in Table 36. The power of SIBTEST increased as test length increased for sample size ratio of 1:1 but its power decreased slightly as test length increased for the ratio of 1:2.

Table 36: Detection Rates for the Test Length by Sample Size Ratio Interaction for the Uniform-DIF (Constant) Condition for the SIBTEST Procedure ((R-$N(0,1)$, F-$N(0,1)$)

	Sample Size Ratio	
Test Length	1:1 Ratio	1:2 Ratio
20-item test	.865	.874
40-item test	.918	.867

Note. Each entry is averaged over 3 sample sizes and 3 item discriminations and is based on 900 observations.

The Type I error rates and power of SIBTEST in detecting uniform DIF (constant) for the condition of equal group abilities are presented in Table 37. Although the Type I error rates of SIBTEST were slightly inflated with a range of .05 to .09 for the 20-item test and .03 to .07 for the 40-item test, they were still within the acceptable range, <.10.

Table 37: The Type I Error and Power of the SIBTEST Procedure in Detecting Uniform DIF (Constant) for the Condition of Equal Group Ability Distributions (R-$N(0,1)$, F-$N(0.1)$)

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.05	0.07	0.05		0.51	0.74	0.90
	600/600		0.07	0.08	0.08		0.76	0.94	0.99
	1200/1200		0.08	0.06	0.06		0.94	1.00	1.00
1:2 Ratio	200/400		0.05	0.06	0.07		0.54	0.76	0.96
	400/800		0.06	0.07	0.05		0.73	0.96	0.99
	800/1600		0.08	0.08	0.09		0.94	0.99	1.00
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.06	0.07	0.05		0.59	0.80	0.98
	600/600		0.04	0.03	0.07		0.89	1.00	1.00
	1200/1200		0.03	0.07	0.04		1.00	1.00	1.00
1:2 Ratio	200/400		0.07	0.06	0.07		0.50	0.73	0.95
	400/800		0.03	0.04	0.04		0.68	0.99	1.00
	800/1600		0.05	0.07	0.04		0.95	1.00	1.00

The power of SIBTEST was unsatisfactory only with the combination of small sample size and low discrimination (<.70); it had adequate to extremely high power in detecting constant uniform DIF in other conditions (with one exception).

Unequal Group Ability Distributions **(R-*N*(0,1), F-*N*(-.5,1))**

When there was a small group ability difference, Type I error of SIBTEST was acceptable when item discrimination was low to medium except when sample size was large and test length was short. Most of the Type I error rates at a=1.5 exceeded the acceptable rate and the power under this condition was not credible; therefore, the outcome under this condition was not included in the logit analysis. Only two levels of item discrimination (a=0.5 and a=1.0) were used in the logit analysis for the condition of small group ability difference.

The result of the logit analysis indicated that for Type I error, a model with only the dependent variable was sufficient to explain the distribution of frequencies observed for SIBTEST (χ^2=15.97, df=23, p=.857). Therefore, none of the independent variables had any effect on its Type I error rates.

For power, the result of the logit analysis indicated that a model with the effects of A, S, R, and AxS was necessary to explain the observed frequencies for the dependent variable for SIBTEST (χ^2=19.04, df=17, p=.326). The detection rates for this interaction are presented in Table 38 and the difference between the two levels of item discrimination decreased as sample size increased. This interaction was due to the ceiling effect observed as sample size and item discrimination increased.

Table 38: Detection Rates for the Item Discrimination by Sample Size Interaction for the Uniform-DIF (Constant) Condition for the SIBTEST Procedure ((R-$N(0,1)$, F-$N(-.5,1)$))

	Sample Size		
Item Discrimination	300/300 200/400	600/600 400/800	1200/1200 800/1600
a=0.5	.563	.790	.913
a=1.0	.753	.948	.983

Note. Each entry is averaged over 2 test lengths and 2 sample size ratios and is based on 400 observations

The Type I error rates and power of SIBTEST in detecting uniform DIF (constant) for the condition of small group ability difference are presented in Table 39. For the main effects of A, S, and R, the power of SIBTEST generally increased as sample size and item discrimination increased. Although sample size ratio was included in the logit model, it was insignificant. It was only with the combination of small sample size and

Table 39: The Type I Error and Power of the SIBTEST Procedure in Detecting Uniform DIF (Constant) for the Condition of Unequal Group Ability Distributions (R-$N(0,1)$, F-$N(-.5,1)$)

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.05	0.07	**0.11**		0.49	0.79	**0.84**
	600/600		0.10	0.09	**0.18**		0.82	0.93	**0.96**
	1200/1200		0.11	0.14	**0.19**		0.92	1.00	**1.00**
1:2 Ratio	200/400		0.05	0.07	**0.09**		0.57	0.78	**0.92**
	400/800		0.07	0.06	**0.15**		0.80	0.94	**0.99**
	800/1600		0.08	0.11	**0.21**		0.95	0.99	**1.00**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.06	0.06	**0.13**		0.63	0.72	**0.93**
	600/600		0.05	0.07	**0.06**		0.76	0.99	**0.99**
	1200/1200		0.08	0.03	**0.13**		0.96	1.00	**1.00**
1:2 Ratio	200/400		0.08	0.09	**0.10**		0.56	0.87	**0.96**
	400/800		0.05	0.07	**0.12**		0.63	0.93	**0.99**
	800/1600		0.08	0.08	**0.10**		0.93	1.00	**1.00**

low discrimination that SIBTEST had a lack of power (with rates ranging from .49 to .63). SIBTEST could detect constant uniform DIF with extremely high power at medium to large sample size and medium to high discrimination. It should be noted that Type I error of SIBTEST was unacceptable when item discrimination was high or when sample size was large. One should not put any faith in its high power under these conditions.

Unequal Group Ability Distributions **(R-*N*(0,1), F-*N*(-1,1))**

The Type I error rates and power of SIBTEST in detecting uniform DIF (constant) for the condition of large group ability difference are presented in Table 40.

Table 40: The Type I Error and Power of the SIBTEST Procedure in Detecting Uniform DIF (Constant) for the Condition of Unequal Group Ability Distributions (R-*N*(0,1), F-*N*(-1,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.05	**0.11**	**0.27**		0.53	**0.80**	**0.90**
	600/600		0.05	**0.27**	**0.35**		0.68	**0.90**	**0.98**
	1200/1200		0.07	**0.35**	**0.42**		0.79	**0.96**	**1.00**
1:2 Ratio	200/400		0.05	**0.12**	**0.30**		0.80	**0.93**	**0.94**
	400/800		0.08	**0.30**	**0.44**		0.78	**0.87**	**0.99**
	800/1600		0.06	**0.41**	**0.47**		0.92	**0.97**	**1.00**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.07	**0.11**	**0.18**		0.68	**0.86**	**0.94**
	600/600		0.05	**0.09**	**0.12**		0.70	**0.92**	**1.00**
	1200/1200		0.07	**0.15**	**0.12**		0.94	**0.99**	**1.00**
1:2 Ratio	200/400		0.06	**0.12**	**0.21**		0.84	**0.96**	**1.00**
	400/800		0.07	**0.14**	**0.27**		0.79	**0.93**	**0.98**
	800/1600		0.09	**0.17**	**0.23**		0.78	**0.96**	**1.00**

As can be seen from Table 40, when the focal group ability was substantially lower than that of the reference group, the Type I error rates of SIBTEST increased

substantially. Its Type I error was acceptable when a=0.5. However, its Type I error was highly inflated (>.10) when a=1.0 or 1.5 (with rates as high as .47) although the inflation was less excessive with the longer test. The outcome at a=1.0 and a=1.5 was not included in the logit analysis because it would be meaningless to further assess the effects of other factors given the unacceptably high Type I error rates for these two levels of item discrimination. Therefore, the logit analysis was conducted for the outcome only at a=0.5.

The result of the logit analysis indicated that for Type I error, a model with only the dependent variable was sufficient to explain the distribution of frequencies observed for the outcome for SIBTEST (χ^2=3.10, df=11, p=.989). Neither test length nor sample size nor sample size ratio had any effect on the Type I error of SIBTEST.

For power, the result of the logit analysis indicated that a model with only the main effect of sample size was adequate (χ^2=7.49, df=9, p=.586). Higher power was associated with larger sample size.

Uniform DIF (Balanced)

The detection rates of the SIBTEST procedure for the uniform-DIF (balanced) condition for each combination of ability distribution difference, test length, item discrimination, sample size, and sample size ratio are presented in Table 41 along with its Type I error rates.

Table 41

The Rejection Rates of the SIBTEST Procedure for the Uniform DIF (Balanced) Condition

		Type I Error			**Power**		
R-N(0,1), F-N(0,1)							
20-item Test		Item Discrimination			Item Discrimination		
Sample Size		a=0.5	a=1.0	a=1.5	a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300	0.05	0.07	0.05	0.06	0.08	0.08
	600/600	0.07	0.08	0.08	0.08	0.09	0.10
	1200/1200	0.08	0.06	0.06	0.12	0.09	0.14
1:2 Ratio	200/400	0.05	0.06	0.07	0.07	0.04	0.06
	400/800	0.06	0.07	0.05	0.08	0.07	0.09
	800/1600	0.08	0.08	0.09	0.10	0.11	0.14
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.06	0.07	0.05	0.07	0.06	0.02
	600/600	0.04	0.03	0.07	0.05	0.06	0.04
	1200/1200	0.03	0.07	0.04	0.02	0.09	0.04
1:2 Ratio	200/400	0.07	0.06	0.07	0.09	0.03	0.08
	400/800	0.03	0.04	0.04	0.06	0.09	0.07
	800/1600	0.05	0.07	0.04	0.06	0.04	0.05
R-N(0,1), F-N(-.5,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.05	0.07	0.11	0.06	0.09	0.15
	600/600	0.10	0.09	0.18	0.08	0.17	0.22
	1200/1200	0.11	0.14	0.19	0.11	0.22	0.34
1:2 Ratio	200/400	0.05	0.07	0.09	0.08	0.08	0.24
	400/800	0.07	0.06	0.15	0.10	0.19	0.24
	800/1600	0.08	0.11	0.21	0.09	0.25	0.31
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.06	0.06	0.13	0.11	0.09	0.11
	600/600	0.05	0.07	0.06	0.01	0.11	0.06
	1200/1200	0.08	0.03	0.13	0.04	0.07	0.10
1:2 Ratio	200/400	0.08	0.09	0.10	0.16	0.19	0.21
	400/800	0.05	0.07	0.12	0.06	0.11	0.15
	800/1600	0.08	0.08	0.10	0.05	0.11	0.16
R-N(0,1), F-N(-1,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.05	0.11	0.27	0.10	0.31	0.37
	600/600	0.05	0.27	0.35	0.13	0.35	0.41
	1200/1200	0.07	0.35	0.42	0.16	0.43	0.43
1:2 Ratio	200/400	0.05	0.12	0.30	0.09	0.45	0.45
	400/800	0.08	0.30	0.44	0.11	0.39	0.43
	800/1600	0.06	0.41	0.47	0.16	0.42	0.44
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.07	0.11	0.18	0.05	0.27	0.50
	600/600	0.05	0.09	0.12	0.08	0.14	0.21
	1200/1200	0.07	0.15	0.12	0.07	0.14	0.12
1:2 Ratio	200/400	0.06	0.12	0.21	0.12	0.69	0.61
	400/800	0.07	0.14	0.27	0.14	0.28	0.37
	800/1600	0.09	0.17	0.23	0.10	0.17	0.15

From Table 41, one concludes that SIBTEST was not useful for detecting uniform DIF of balanced type. When two groups had equal abilities, SIBTEST failed to detect items with uniform DIF of balanced type almost completely. However, like its Type I error rates, its detection rates increased substantially as the between-group ability differences increased. It probably detected group ability difference or impact rather than DIF. No logit analysis was conducted since the detection rates of SIBTEST for balanced uniform DIF were reflections of its Type I error rates.

Nonuniform DIF

The detection rates of the SIBTEST procedure for the nonuniform DIF condition for each combination of ability distribution difference, test length, item discrimination, sample size, and sample size ratio are presented in Table 42 along with its Type I error rates.

As in the uniform-DIF (balanced) condition, SIBTEST was not useful for detecting nonuniform DIF. This was expected since SIBTEST is not designed to detect nonuniform DIF. No logit analysis was conducted because these detection rates obviously reflected the Type I error rates of SIBTEST.

In summary, the Type I error rates of SIBTEST were slightly inflated but still acceptable when groups were equal in ability or when there was a small group ability difference but item discrimination was low to medium (with a couple of exceptions) or when there was a large group ability difference and item discrimination was low; under the above conditions, SIBTEST could be used in detecting constant uniform DIF with

Table 42

The Rejection Rates of the SIBTEST Procedure for the Nonuniform DIF Condition

		Type I Error			Power		
R-$N(0,1)$, F-$N(0,1)$							
20-item Test		Item Discrimination			a-Parameter Difference		
Sample		a=0.5	a=1.0	a=1.5	a_D=0.5	a_D=1.0	a_D=1.5
	300/300	0.05	0.07	0.05	0.07	0.05	0.08
	600/600	0.07	0.08	0.08	0.09	0.10	0.11
	1200/1200	0.08	0.06	0.06	0.12	0.10	0.13
	200/400	0.05	0.06	0.07	0.08	0.03	0.08
	400/800	0.06	0.07	0.05	0.11	0.07	0.12
	800/1600	0.08	0.08	0.09	0.10	0.09	0.11
40-item Test							
Sample							
	300/300	0.06	0.07	0.05	0.10	0.04	0.06
	600/600	0.04	0.03	0.07	0.08	0.07	0.08
	1200/1200	0.03	0.07	0.04	0.08	0.11	0.08
	200/400	0.07	0.06	0.07	0.03	0.06	0.05
	400/800	0.03	0.04	0.04	0.07	0.07	0.08
	800/1600	0.05	0.07	0.04	0.13	0.07	0.17
R-$N(0,1)$, F-$N(-.5,1)$							
20-item Test							
Sample							
	300/300	0.05	0.07	0.11	0.06	0.15	0.16
	600/600	0.10	0.09	0.18	0.06	0.24	0.27
	1200/1200	0.11	0.14	0.19	0.12	0.39	0.40
	200/400	0.05	0.07	0.09	0.08	0.16	0.18
	400/800	0.07	0.06	0.15	0.16	0.19	0.22
	800/1600	0.08	0.11	0.21	0.10	0.23	0.25
40-item Test							
Sample							
	300/300	0.06	0.06	0.13	0.10	0.13	0.12
	600/600	0.05	0.07	0.06	0.14	0.20	0.27
	1200/1200	0.08	0.03	0.13	0.18	0.31	0.36
	200/400	0.08	0.09	0.10	0.12	0.13	0.15
	400/800	0.05	0.07	0.12	0.16	0.17	0.20
	800/1600	0.08	0.08	0.10	0.33	0.40	0.45
R-$N(0,1)$, F-$N(-1,1)$							
20-item Test							
Sample							
	300/300	0.05	0.11	0.27	0.10	0.17	0.13
	600/600	0.05	0.27	0.35	0.08	0.14	0.25
	1200/1200	0.07	0.35	0.42	0.23	0.26	0.33
	200/400	0.05	0.12	0.30	0.11	0.15	0.25
	400/800	0.08	0.30	0.44	0.14	0.21	0.26
	800/1600	0.06	0.41	0.47	0.29	0.35	0.40
40-item Test							
Sample							
	300/300	0.07	0.11	0.18	0.14	0.15	0.27
	600/600	0.05	0.09	0.12	0.27	0.40	0.66
	1200/1200	0.07	0.15	0.12	0.50	0.60	0.73
	200/400	0.06	0.12	0.21	0.23	0.24	0.25
	400/800	0.07	0.14	0.27	0.25	0.25	0.36
	800/1600	0.09	0.17	0.23	0.43	0.62	0.72

adequate to extremely high power except when item discrimination was low and sample size was small. SIBTEST should not be used with the combination of small group ability difference and high discrimination and the combination of large group ability difference and medium to high discrimination because its Type I error was in general unacceptable under these conditions. SIBTEST was not useful in detecting balanced uniform DIF or nonuniform DIF and its detection rates for these two types of DIF were very similar to its Type I error rates.

The LDFA-Uniform Procedure

With the LDFA procedure, the hypotheses of no uniform DIF and nonuniform DIF were tested separately. Results for LDFA-Uniform and LDFA-Nonuniform test statistics are reported separately.

Uniform DIF (Constant)

Equal Group Ability Distributions **(R-*N*(0,1), F-*N*(0,1))**

The result of logit analysis indicated that for Type I error, a model with only the dependent variable was sufficient to explain the distribution of frequencies observed for LDFA-Uniform (χ^2=17.38, df=35, p=.994). None of the independent variables had any effect on the Type I error rates of LDFA-Uniform.

For power, the result of logit analysis indicated that a model with the effects of A, S, and AxS was needed to sufficiently explain the frequencies of the outcome variable (χ^2= 32.81, df=27, p=.203). The detection rates for the AxS interaction for LDFA-

Uniform under the condition of equal group abilities are presented in Table 43. As can be seen from the table, the differences among the three levels of item discrimination decreased as sample size increased; similarly, the differences across sample sizes decreased as item discrimination increased. This interaction is really due to the ceiling effect observed as discrimination and sample size increased.

Table 43: Detection Rates for the Item Discrimination by Sample Size Interaction for the Uniform-DIF (Constant) Condition for the LDFA-Uniform Procedure (R-$N(0,1)$, F-$N(0,1)$)

	Sample Size		
Item Discrimination	300/300 200/400	600/600 400/800	1200/1200 800/1600
a=0.5	.560	.785	.968
a=1.0	.795	.990	1.00
a=1.5	.963	.998	1.00

Note. Each entry is averaged over 2 test lengths and 2 sample size ratios is based on 400 observations.

The Type I error rates and power of LDFA-Uniform in detecting uniform DIF (constant) for the condition of equal group abilities are presented in Table 44. As can be seen from Table 44, the Type I error rates of LDFA-Uniform were within or close to .05 in most conditions.

Since LDFA-Uniform had excellent control over Type I error when two groups were equal in ability, its power can be assessed with confidence under this condition. Based on the criteria set in this study, LDFA-Uniform was extremely powerful in detecting constant uniform DIF with the combination of medium to large sample size and medium to high discrimination, the combination of small sample size and high

Table 44: The Type I Error and Power of the LDFA-Uniform Procedure in Detecting Uniform DIF (Constant) for the Condition of Equal Group Ability Distributions (R-$N(0,1)$, F-$N(0,1)$)

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.04	0.05	0.03		0.57	0.76	0.96
	600/600		0.06	0.05	0.07		0.77	0.98	0.99
	1200/1200		0.06	0.03	0.03		0.96	1.00	1.00
1:2 Ratio	200/400		0.04	0.03	0.06		0.51	0.73	0.96
	400/800		0.05	0.05	0.03		0.75	0.99	1.00
	800/1600		0.07	0.05	0.07		0.97	1.00	1.00
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.04	0.07	0.04		0.63	0.86	0.97
	600/600		0.02	0.04	0.06		0.88	1.00	1.00
	1200/1200		0.03	0.07	0.04		1.00	1.00	1.00
1:2 Ratio	200/400		0.07	0.06	0.05		0.53	0.83	0.96
	400/800		0.03	0.05	0.04		0.74	0.99	1.00
	800/1600		0.06	0.07	0.04		0.94	1.00	1.00

discrimination, and the combination of large sample size and low discrimination. Under these conditions, the detection rates ranged from .94 to 1.0. The power of LDFA-Uniform was adequate with the combination of small sample size and medium discrimination and the combination of medium sample size and low discrimination. LDFA-Uniform lacked power in detecting constant uniform DIF only with the combination of low discrimination and small sample size.

Unequal Group Ability Distributions **(R-$N(0,1)$, F-$N(-.5,1)$)**

The Type I error rates and power of LDFA-Uniform in detecting uniform DIF (constant) for the condition of small group ability difference are presented in Table 45. When there was a small group ability difference, the Type I error rates of LDFA-Uniform increased considerably. In general, its Type I error was in good control when item

Table 45: The Type I Error and Power of the LDFA-Uniform Procedure in Detecting Uniform DIF (Constant) for the Condition of Unequal Group Ability Distributions (R-$N(0,1)$, F-$N(-.5,1)$)

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.03	0.08	**0.19**		0.47	0.92	**0.99**
	600/600		0.02	0.07	**0.18**		0.86	0.97	**1.00**
	1200/1200		0.07	0.18	**0.40**		0.95	1.00	**1.00**
1:2 Ratio	200/400		0.03	0.09	**0.15**		0.36	0.84	**0.96**
	400/800		0.02	0.06	**0.25**		0.76	0.97	**1.00**
	800/1600		0.06	0.16	**0.45**		0.95	1.00	**1.00**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.04	0.07	**0.12**		0.57	0.89	**0.98**
	600/600		0.05	0.06	**0.10**		0.80	1.00	**1.00**
	1200/1200		0.09	0.09	**0.16**		0.98	1.00	**1.00**
1:2 Ratio	200/400		0.04	0.07	**0.08**		0.43	0.92	**0.98**
	400/800		0.02	0.10	**0.14**		0.66	1.00	**1.00**
	800/1600		0.07	0.06	**0.12**		0.95	1.00	**1.00**

discrimination was low. Type I error was inflated but acceptable (<.10) when item discrimination was medium except when sample size was large and test length was short. Its Type I error exceeded .10 when item discrimination was high and the power under this condition was questionable. The outcomes under this condition were not included in the logit analysis. Only two levels of item discrimination (a=0.5 and a=1.0) were used in the logit analysis for the condition of small group ability difference.

The result of the logit analysis indicated that for Type I error, a model with the main effects of A and S was adequate to explain the frequency distribution of the dependent variable for LDFA-Uniform (χ^2=15.61, df=20, p=.741). Increasing sample size and item discrimination led to an increase in Type I error rates.

For power, the result of logit analysis indicated that a model with the effects of A, S, R, and AxS was necessary to explain the frequencies of the outcome variable for LDFA-Uniform (χ^2=21.72, df=17, p=.196). The detection rates for the AxS interaction for LDFA-Uniform for the condition of small group ability difference are presented in Table 46. As can be seen from the table, the difference between the two levels of item discrimination decreased as sample size increased; similarly, the difference across sample sizes decreased as item discrimination increased. This interaction is basically due to the ceiling effect observed as item discrimination and sample size increased.

Table 46: Detection Rates for the Item Discrimination by Sample Size Interaction for the Uniform-DIF (Constant) Condition for the LDFA-Uniform Procedure (R-N(0,1), F-N(0,1))

	Sample Size		
Item Discrimination	300/300 200/400	600/600 400/800	1200/1200 800/1600
a=0.5	.458	.770	.958
a=1.0	.893	.985	1.00

Note. Each entry is averaged over 2 test lengths and 2 sample size ratios is based on 400 observations.

For the main effects, the power of LDFA-Uniform increased as sample size and item discrimination increased, but its power decreased slightly when the sample size ratio changed from 1:1 to 1:2. Except for the combination of low discrimination and small sample size, LDFA-Uniform could detect constant uniform DIF with adequate to perfect accuracy under all other conditions. It should be noted that the Type I error of LDFA-Uniform was unacceptably high when item discrimination was high and sample size was

large. LDFA-Uniform's high detection rates for uniform DIF (constant) under these conditions were due to both high power and high Type I error rates.

Unequal Group Ability Distributions **(R-*N*(0,1), F-*N*(-1,1))**

The Type I error rates and power of LDFA-Uniform in detecting uniform DIF (constant) for the condition of large group ability difference are presented in Table 47.

Table 47: The Type I Error and Power of the LDFA-Uniform Procedure in Detecting Uniform DIF (Constant) for the Condition of Unequal Group Ability Distributions (R-*N*(0,1), F-*N*(-1,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.01	**0.14**	**0.37**		0.42	**0.94**	**1.00**
	600/600		0.03	**0.33**	**0.69**		0.73	**1.00**	**1.00**
	1200/1200		0.04	**0.50**	**0.98**		0.99	**1.00**	**1.00**
1:2 Ratio	200/400		0.02	**0.18**	**0.36**		0.48	**0.85**	**1.00**
	400/800		0.04	**0.36**	**0.67**		0.73	**1.00**	**1.00**
	800/1600		0.10	**0.55**	**0.93**		0.95	**1.00**	**1.00**
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.03	**0.06**	**0.18**		0.59	**0.87**	**0.99**
	600/600		0.05	**0.07**	**0.30**		0.75	**1.00**	**1.00**
	1200/1200		0.06	**0.23**	**0.54**		0.98	**1.00**	**1.00**
1:2 Ratio	200/400		0.04	**0.07**	**0.09**		0.46	**0.90**	**1.00**
	400/800		0.04	**0.13**	**0.32**		0.77	**0.98**	**1.00**
	800/1600		0.06	**0.17**	**0.43**		0.95	**1.00**	**1.00**

As can be seen from Table 47, when the focal group ability was substantially lower than that of the reference group, the Type I error rates of LDFA-Uniform increased substantially. Its Type I error was in good control when a=0.5. However, its Type I error was highly inflated (>.10) when a=1.0 or 1.5 (with rates as high as .98) although the inflation was less excessive with the longer test. Like all the other procedures, LDFA-

Uniform detected impact rather than DIF. The outcome at a=1.0 and a=1.5 was not included in the logit analysis because the power was not credible for these two levels of item discrimination due to the excessive Type I error. Therefore, the logit analysis was conducted for the outcome only at a=0.5.

The result of the logit analysis indicated that for Type I error, a model with only the dependent variable was sufficient to explain the distribution of frequencies observed for the outcome for LDFA-Uniform (χ^2=13.58, df=11, p=.257). Neither test length nor sample size nor sample size ratio had any effect on the Type I error of LDFA-Uniform.

For power, a model with the main effect of sample size was sufficient to explain the frequencies of the outcome variable (χ^2= 11.41, df=9, p=.249). The power of LDFA-Uniform at a=0.5 increased with the increase of sample size, however, the power was inadequate when sample size was small.

Uniform DIF (Balanced)

The rejection rates of the LDFA-Uniform procedure for the uniform-DIF (balanced) condition are presented in Table 48 along with its Type I error rates. As shown in Table 48, LDFA-Uniform was not powerful in detecting uniform DIF of balanced type. When two groups had equal abilities, LDFA-Uniform failed to detect items with uniform DIF of balanced type almost completely. However, its detection rates increased as the between-group ability differences increased and so did its Type I error. No logit analysis was conducted since the detection rates of LDFA-Uniform for balanced uniform DIF were very similar to its Type I error rates.

Table 48

The Rejection Rates of LDFA-Uniform for the Uniform DIF (Balanced) Condition

		Type I Error			**Power**		
R-$N(0,1)$, F-$N(0,1)$							
20-item Test		Item Discrimination			Item Discrimination		
Sample Size		a=0.5	a=1.0	a=1.5	a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300	0.04	0.05	0.03	0.05	0.05	0.06
	600/600	0.06	0.05	0.07	0.07	0.06	0.05
	1200/1200	0.06	0.03	0.03	0.06	0.04	0.06
1:2 Ratio	200/400	0.04	0.03	0.06	0.07	0.05	0.13
	400/800	0.05	0.05	0.03	0.00	0.06	0.06
	800/1600	0.07	0.05	0.07	0.04	0.06	0.06
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.04	0.07	0.04	0.07	0.05	0.00
	600/600	0.02	0.04	0.06	0.06	0.07	0.04
	1200/1200	0.03	0.07	0.04	0.04	0.07	0.04
1:2 Ratio	200/400	0.07	0.06	0.05	0.05	0.01	0.07
	400/800	0.03	0.05	0.04	0.06	0.09	0.05
	800/1600	0.06	0.07	0.04	0.06	0.03	0.04
R-$N(0,1)$, F-$N(-.5,1)$							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.03	0.08	0.19	0.09	0.09	0.25
	600/600	0.02	0.07	0.18	0.05	0.20	0.44
	1200/1200	0.07	0.18	0.40	0.07	0.22	0.71
1:2 Ratio	200/400	0.03	0.09	0.15	0.03	0.10	0.28
	400/800	0.02	0.06	0.25	0.06	0.23	0.46
	800/1600	0.06	0.16	0.45	0.05	0.29	0.65
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.04	0.07	0.12	0.08	0.07	0.09
	600/600	0.05	0.06	0.10	0.04	0.13	0.27
	1200/1200	0.09	0.09	0.16	0.03	0.19	0.40
1:2 Ratio	200/400	0.04	0.07	0.08	0.07	0.16	0.18
	400/800	0.02	0.10	0.14	0.04	0.14	0.28
	800/1600	0.07	0.06	0.12	0.05	0.28	0.51
R-$N(0,1)$, F-$N(-1,1)$							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.01	0.14	0.37	0.09	0.31	0.75
	600/600	0.03	0.33	0.69	0.14	0.49	0.94
	1200/1200	0.04	0.50	0.98	0.10	0.84	1.00
1:2 Ratio	200/400	0.02	0.18	0.36	0.10	0.40	0.65
	400/800	0.04	0.36	0.67	0.07	0.66	0.95
	800/1600	0.10	0.55	0.93	0.13	0.94	1.00
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.03	0.06	0.18	0.07	0.15	0.51
	600/600	0.05	0.07	0.30	0.06	0.31	0.69
	1200/1200	0.06	0.23	0.54	0.05	0.56	0.96
1:2 Ratio	200/400	0.04	0.07	0.09	0.09	0.24	0.42
	400/800	0.04	0.13	0.32	0.12	0.34	0.79
	800/1600	0.06	0.17	0.43	0.14	0.51	0.95

Nonuniform DIF

The rejection rates of the LDFA-Uniform procedure for the nonuniform condition are presented in Table 49 along with its Type I error rates.

As in the uniform-DIF (balanced) condition, the detection rates of LDFA-Uniform for the nonuniform-DIF condition were very similar to its Type I error rates. When the reference and focal groups had equal ability distributions, the detection rates of LDFA-Uniform were close to zero. This was expected since LDFA-Uniform is not designed to detect nonuniform DIF. However, like its Type I error, its power increased as the between-group ability difference increased. No logit analysis was conducted.

In summary, LDFA-Uniform had very good control over Type I error when groups were equal in ability or when groups differed in ability but item discrimination was low. Its Type I error was inflated but still acceptable when there was a small group ability difference and item discrimination was medium (except for large sample size and shorter test length); under the above conditions, LDFA-Uniform could be used in detecting constant uniform DIF with adequate to extremely high power even with small sample size (except when item discrimination was low). LDFA-Uniform should not be used with the combination of small group ability difference and high discrimination and the combination of large group ability difference and medium to high discrimination because its Type I error was unacceptably high under these conditions. LDFA-Uniform was not useful in detecting balanced uniform DIF or nonuniform DIF and its detection rates for these two types of DIF were very similar to its Type I error rates.

Table 49

The Rejection Rates of the LDFA-Uniform Procedure for the Nonuniform DIF Condition

		Type I Error			Power		
R-N(0,1), F-N(0,1)							
20-item Test		Item Discrimination			a-Parameter Difference		
Sample Size		a=0.5	a=1.0	a=1.5	a_D=0.5	a_D=1.0	a_D=1.5
1:1 Ratio	300/300	0.04	0.05	0.03	0.05	0.01	0.04
	600/600	0.06	0.05	0.07	0.04	0.01	0.05
	1200/1200	0.06	0.03	0.03	0.02	0.02	0.04
1:2 Ratio	200/400	0.04	0.03	0.06	0.06	0.02	0.07
	400/800	0.05	0.05	0.03	0.06	0.08	0.06
	800/1600	0.07	0.05	0.07	0.09	0.05	0.07
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.04	0.07	0.04	0.06	0.02	0.04
	600/600	0.02	0.04	0.06	0.06	0.04	0.08
	1200/1200	0.03	0.07	0.04	0.05	0.02	0.04
1:2 Ratio	200/400	0.07	0.06	0.05	0.05	0.07	0.09
	400/800	0.03	0.05	0.04	0.06	0.13	0.07
	800/1600	0.06	0.07	0.04	0.08	0.09	0.06
R-N(0,1), F-N(-.5,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.03	0.08	0.19	0.07	0.05	0.10
	600/600	0.02	0.07	0.18	0.04	0.09	0.11
	1200/1200	0.07	0.18	0.40	0.09	0.16	0.10
1:2 Ratio	200/400	0.03	0.09	0.15	0.12	0.19	0.06
	400/800	0.02	0.06	0.25	0.15	0.17	0.18
	800/1600	0.06	0.16	0.45	0.09	0.30	0.16
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.04	0.07	0.12	0.07	0.12	0.09
	600/600	0.05	0.06	0.10	0.11	0.10	0.18
	1200/1200	0.09	0.09	0.16	0.13	0.24	0.34
1:2 Ratio	200/400	0.04	0.07	0.08	0.09	0.18	0.24
	400/800	0.02	0.10	0.14	0.15	0.29	0.32
	800/1600	0.07	0.06	0.12	0.27	0.48	0.49
R-N(0,1), F-N(-1,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.01	0.14	0.37	0.09	0.13	0.15
	600/600	0.03	0.33	0.69	0.08	0.11	0.27
	1200/1200	0.04	0.50	0.98	0.19	0.24	0.35
1:2 Ratio	200/400	0.02	0.18	0.36	0.09	0.20	0.22
	400/800	0.04	0.36	0.67	0.11	0.23	0.29
	800/1600	0.10	0.55	0.93	0.20	0.45	0.48
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.03	0.06	0.18	0.08	0.13	0.22
	600/600	0.05	0.07	0.30	0.17	0.24	0.24
	1200/1200	0.06	0.23	0.54	0.27	0.25	0.27
1:2 Ratio	200/400	0.04	0.07	0.09	0.22	0.33	0.49
	400/800	0.04	0.13	0.32	0.26	0.46	0.67
	800/1600	0.06	0.17	0.43	0.49	0.81	0.87

The LDFA-Nonuniform Procedure

Uniform DIF (Constant)

The Type I error rates and detection rates of the LDFA-Nonuniform procedure for the uniform DIF (constant) condition are reported in Table 50.

Equal Group Ability Distributions **(R-*N*(0,1), F-*N*(0,1))**

The result of logit analysis indicated that for Type I error, a model with only the dependent variable was sufficient to explain the distribution of frequencies observed for LDFA-Nonuniform (χ^2=20.36, df=35, p=.977). None of the independent variables had any effect on the Type I error rates of LDFA-Nonuniform.

As can be seen from Table 50, LDFA-Nonuniform was unable to detect items with constant uniform DIF. Since no group difference in item discrimination was introduced into the studied items, the detection rates for LDFA-Nonuniform were quite comparable to its Type I error rates. No logit analysis was conducted since the detection rates of LDFA-Nonuniform for all three levels of group ability difference were actually its Type I error rates.

Unequal Group Ability Distributions **(R-*N*(0,1), F-*N*(-.5,1))**

When there was a small group ability difference, the Type I error rates of LDFA-Nonuniform were generally within the acceptable range (<.10) with a few exceptions. The results of the logit analysis indicated that for LDFA-Nonuniform, even a model including all three-way interactions as well as all two-way interactions and main effects did not provide adequate fit to the dependent variable. Therefore, a saturated model including the four-way interaction SxRxLxA had to be used. No attempt was made to

Table 50

Type I Error and Power of LDFA-Nonuniform in Detecting Uniform DIF (Constant)

		Type I Error			**Power**		
R-N(0,1), F-N(0,1)							
20-item Test		Item Discrimination			Item Discrimination		
Sample Size		a=0.5	a=1.0	a=1.5	a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300	0.05	0.03	0.05	0.11	0.05	0.05
	600/600	0.04	0.05	0.03	0.04	0.06	0.06
	1200/1200	0.03	0.03	0.06	0.07	0.05	0.09
1:2 Ratio	200/400	0.05	0.06	0.06	0.05	0.03	0.07
	400/800	0.05	0.06	0.07	0.09	0.06	0.04
	800/1600	0.03	0.03	0.05	0.06	0.03	0.03
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.05	0.03	0.06	0.05	0.04	0.10
	600/600	0.04	0.05	0.06	0.07	0.07	0.08
	1200/1200	0.08	0.04	0.07	0.05	0.03	0.04
1:2 Ratio	200/400	0.03	0.05	0.03	0.03	0.03	0.06
	400/800	0.06	0.06	0.07	0.02	0.05	0.02
	800/1600	0.03	0.01	0.04	0.06	0.02	0.04
R-N(0,1), F-N(-.5,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.02	0.05	0.07	0.03	0.03	0.07
	600/600	0.06	0.13	0.05	0.09	0.07	0.04
	1200/1200	0.08	0.07	0.08	0.06	0.04	0.11
1:2 Ratio	200/400	0.00	0.12	0.03	0.04	0.05	0.09
	400/800	0.07	0.12	0.06	0.05	0.02	0.04
	800/1600	0.05	0.03	0.03	0.09	0.05	0.07
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.05	0.08	0.05	0.04	0.08	0.05
	600/600	0.07	0.07	0.10	0.06	0.09	0.09
	1200/1200	0.09	0.08	0.07	0.09	0.09	0.08
1:2 Ratio	200/400	0.13	0.05	0.08	0.04	0.04	0.04
	400/800	0.04	0.08	0.06	0.02	0.05	0.10
	800/1600	0.08	0.10	0.13	0.08	0.08	0.15
R-N(0,1), F-N(-1,1)							
20-item Test							
Sample Size							
1:1 Ratio	300/300	0.09	0.10	0.07	0.06	0.11	0.20
	600/600	0.10	0.13	0.12	0.14	0.17	0.23
	1200/1200	0.16	0.25	0.21	0.17	0.19	0.28
1:2 Ratio	200/400	0.10	0.06	0.13	0.06	0.12	0.13
	400/800	0.12	0.17	0.09	0.09	0.18	0.24
	800/1600	0.16	0.28	0.29	0.24	0.24	0.37
40-item Test							
Sample Size							
1:1 Ratio	300/300	0.10	0.13	0.12	0.09	0.07	0.09
	600/600	0.21	0.23	0.13	0.09	0.15	0.32
	1200/1200	0.27	0.39	0.37	0.22	0.28	0.50
1:2 Ratio	200/400	0.09	0.13	0.12	0.09	0.13	0.18
	400/800	0.17	0.17	0.16	0.13	0.14	0.25
	800/1600	0.26	0.36	0.26	0.18	0.15	0.39

interpret this four-way interaction as it seemed unwieldy. From Table 50, one can see that Type I error rates of LDFA-Nonuniform did not increase or decrease systematically as test length, item discrimination, sample size, and sample size ratio changed.

***Unequal Group Ability Distributions* (R-*N*(0,1), F-*N*(-1,1))**

No logit analysis was carried out for LDFA-Nonuniform for the condition of large group ability difference as its Type I error rates were unacceptably high across all three levels of item discrimination.

Uniform DIF (Balanced)

***Equal Group Ability Distributions* (R-*N*(0,1), F-*N*(0,1))**

The power of LDFA-Nonuniform in detecting uniform-DIF (balanced) for the condition of equal group abilities is presented in Table 51 along with its Type I error rates. The result of the logit analysis indicated that for the power of LDFA-Nonuniform in detecting balanced uniform-DIF, a model with the effects of A and S was adequate to explain the outcome for LDFA-Nonuniform (χ^2=37.19, df=31, p=.205). As can be seen from Table 51, the power of LDFA-Nonuniform in detecting balanced uniform DIF was very low in most conditions. Even when sample size was large and item discrimination was high, its power was borderline and only ranged from .66 to .79. In general, its power increased as sample size and item discrimination increased.

***Unequal Group Ability Distributions* (R-*N*(0,1), F-*N*(-.5,1))**

When there was a small group ability difference, unlike other procedures, Type I error rates of LDFA-Nonuniform were such that it was possible to examine power at all

Table 51: The Type I Error and Power of the LDFA-Nonuniform Procedure in Detecting Uniform DIF (Balanced) for the Condition of Equal Group Ability Distributions (R-$N(0,1)$, F-$N(0,1)$)

			Type I Error				Power		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.05	0.03	0.05		0.09	0.18	0.23
	600/600		0.04	0.05	0.03		0.15	0.35	0.52
	1200/1200		0.03	0.03	0.06		0.24	0.58	0.79
1:2 Ratio	200/400		0.05	0.06	0.06		0.05	0.19	0.25
	400/800		0.05	0.06	0.07		0.15	0.34	0.42
	800/1600		0.03	0.03	0.05		0.24	0.51	0.66
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.05	0.03	0.06		0.10	0.14	0.20
	600/600		0.04	0.05	0.06		0.10	0.29	0.51
	1200/1200		0.08	0.04	0.07		0.27	0.54	0.69
1:2 Ratio	200/400		0.03	0.05	0.03		0.09	0.17	0.14
	400/800		0.06	0.06	0.07		0.05	0.20	0.28
	800/1600		0.03	0.01	0.04		0.20	0.47	0.72

three levels of item discrimination. The result of logit analysis indicated that a model with the effects of A, S, L, and AxS was needed to explain the outcome for LDFA-Nonuniform (χ^2=31.92, df=26, p=.196). The detection rates for the AxS interaction are presented in Table 52. As can be seen from the table, the difference between a=0.5 and a=1 increased as sample size increased but the difference between a=1 and a=1.5 decreased as sample size increased.

The power of LDFA-Nonuniform in detecting uniform-DIF (balanced) for the condition of small group ability difference is presented in Table 53 along with its Type I error rates. LDFA-Nonuniform only had high power when sample size was large and item discrimination was high; its power was borderline (around .70) when sample size was large and item discrimination was medium; its power was inadequate and even very low

Table 52: Detection Rates for the Item Discrimination by Sample Size Interaction for the Uniform-DIF (Balanced) Condition for the LDFA-Nonuniform Procedure (R-N(0,1), F-N(-.5,1))

	Sample Size		
Item Discrimination	300/300 200/400	600/600 400/800	1200/1200 800/1600
a=0.5	.095	.208	.333
a=1.0	.225	.380	.663
a=1.5	.348	.535	.86

Note. Each entry is averaged over 2 test lengths and 2 sample size ratios and is based on 400 observations.

Table 53: The Type I Error and Power of the LDFA-Nonuniform Procedure in Detecting Uniform DIF (Balanced) for the Condition of Unequal Group Ability Distributions (R-N(0,1), F-N(-.5,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.02	0.05	0.07		0.10	0.21	0.44
	600/600		0.06	0.13	0.05		0.23	0.39	0.63
	1200/1200		0.08	0.07	0.08		0.37	0.65	0.90
1:2 Ratio	200/400		0.00	0.12	0.03		0.09	0.26	0.29
	400/800		0.07	0.12	0.06		0.18	0.44	0.47
	800/1600		0.05	0.03	0.03		0.38	0.71	0.85
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.05	0.08	0.05		0.06	0.22	0.37
	600/600		0.07	0.07	0.10		0.25	0.34	0.49
	1200/1200		0.09	0.08	0.07		0.29	0.73	0.88
1:2 Ratio	200/400		0.13	0.05	0.08		0.13	0.21	0.29
	400/800		0.04	0.08	0.06		0.17	0.35	0.55
	800/1600		0.08	0.10	0.13		0.29	0.56	0.81

in all other conditions. In general, its power increased as sample size and item discrimination increased but decreased as test length increased.

Unequal Group Ability Distributions **(R-*N*(0,1), F-*N*(-1,1))**

The power of LDFA-Nonuniform in detecting uniform-DIF (balanced) for the condition of large group ability difference are presented in Table 54 along with its Type I error rates. No logit analysis was conducted for the case of large group ability difference as the Type I error of LDFA-Nonuniform was high in most conditions.

Table 54: The Type I Error and Power of the LDFA-Nonuniform Procedure in Detecting Uniform DIF (Balanced) for the Condition of Unequal Group Ability Distributions (R-*N*(0,1), F-*N*(-1,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				Item Discrimination		
Sample Size			a=0.5	a=1.0	a=1.5		a=0.5	a=1.0	a=1.5
1:1 Ratio	300/300		0.09	0.10	0.07		0.18	0.32	0.48
	600/600		0.10	0.13	0.12		0.38	0.61	0.68
	1200/1200		0.16	0.25	0.21		0.53	0.88	0.95
1:2 Ratio	200/400		0.10	0.06	0.13		0.19	0.31	0.44
	400/800		0.12	0.17	0.09		0.27	0.54	0.67
	800/1600		0.16	0.28	0.29		0.46	0.79	0.94
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.10	0.13	0.12		0.23	0.40	0.46
	600/600		0.21	0.23	0.13		0.35	0.58	0.75
	1200/1200		0.27	0.39	0.37		0.67	0.93	0.98
1:2 Ratio	200/400		0.09	0.13	0.12		0.19	0.28	0.36
	400/800		0.17	0.17	0.16		0.33	0.60	0.66
	800/1600		0.26	0.36	0.26		0.55	0.81	0.82

Nonuniform DIF

Equal Group Ability Distributions **(R-*N*(0,1), F-*N*(0,1))**

The result of the logit analysis indicated that a model including A_D and S main effects was adequate to explain the variations observed in the outcome (χ^2=33.56, df=31, p=.34). The detection rates of LDFA-Nonuniform in detecting nonuniform DIF for the

condition of equal group ability are presented in Table 55 along with its Type I error rates.

Table 55: The Type I Error and Power of the LDFA-Nonuniform Procedure in Detecting Nonuniform DIF for the Condition of Equal Group Ability Distributions (R-$N(0,1)$, F-$N(0,1)$)

			Type I Error				Power		
20-item Test			Item Discrimination				a-Parameter Difference		
Sample Size			a=0.5	a=1.0	a=1.5		a_D=0.5	a_D=1.0	a_D=1.5
1:1 Ratio	300/300		0.05	0.03	0.05		0.28	0.58	0.57
	600/600		0.04	0.05	0.03		0.47	0.70	0.79
	1200/1200		0.03	0.03	0.06		0.77	0.98	0.99
1:2 Ratio	200/400		0.05	0.06	0.06		0.23	0.47	0.53
	400/800		0.05	0.06	0.07		0.47	0.71	0.78
	800/1600		0.03	0.03	0.05		0.77	0.97	0.98
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.05	0.03	0.06		0.38	0.65	0.76
	600/600		0.04	0.05	0.06		0.55	0.89	0.96
	1200/1200		0.08	0.04	0.07		0.96	1.00	0.99
1:2 Ratio	200/400		0.03	0.05	0.03		0.29	0.60	0.78
	400/800		0.06	0.06	0.07		0.71	0.93	0.92
	800/1600		0.03	0.01	0.04		0.92	0.99	1.00

From Table 55, one can see that the power of LDFA-Nonuniform increased as sample size and a-parameter difference increased. The power of LDFA-Nonuniform was adequate to extremely high in detecting nonuniform DIF with the combination of large sample size and medium to high discrimination, the combination of large sample size and low discrimination, and the combination of medium sample size and medium to high discrimination (with power ranging from .70 to 1.0). Since LDFA-Nonuniform had very good control over Type I error when there was no group ability difference, one can have confidence in its high power for the above conditions. However, it had inadequate power

under other conditions. Sample sizes of 300/300 and 200/400 appeared to be inadequate except for the combination of 40-item test, high discrimination, and sample size ratio of 1:1.

Unequal Group Ability Distributions **(R-*N*(0,1), F-*N*(-.5,1))**

When there was a small group ability difference, a model including the interaction LxSxR, A_DxL, A_DxS, A_DxR as well as main effects A_D, L, R, and S was needed to explain the variations observed in the dependent variable (χ^2=14.12, df=14, p=.441). Among the interactions, only A_DxL and A_DxS had standardized parameter estimates greater than $|1.96|$. The detection rates for these two interactions are presented in Table 56. In terms of the A_DxL interaction, the difference between the two test lengths was affected by *a*-parameter difference. The difference between test lengths increased as *a*-parameter difference increased. Regarding the A_DxS interaction, the differences among the three levels of *a*-parameter difference were much smaller for the largest sample size.

Table 56: Detection Rates for the *a*-Parameter Difference by Test Length, *a*-Parameter Difference by Sample Size Interactions for the Nonuniform-DIF Condition for the LDFA-Nonuniform Procedure (R-*N*(0,1), F-*N*(-.5,1))

	Test Length		Sample		
a-Parameter Difference	20-item test	40-item test	300/300 200/400	600/600 400/800	1200/1200 800/1600
a_R-a_F=0.5	.353	.368	.208	.313	.719
a_R-a_F =1.0	.543	.622	.355	.590	.803
a_R-a_F =1.5	.630	.728	.460	.698	.882

Note. For A_DxL interaction, each entry is averaged over 6 sample size combinations and is based on 600 observations; for A_DxS interaction, each entry is averaged over 2 test lengths and 2 sample size ratios and is based on 400 observations.

The power of LDFA-Nonuniform in detecting nonuniform DIF for the condition of small group ability difference is presented in Table 57 along with its Type I error rates. In general, the power of LDFA-Nonuniform increased as test length (except for large sample size and the ratio of 1:2), sample size and item discrimination increased, and its power decreased as sample size ratio changed from 1:1 to 1:2. It is worth noting that the power of LDFA-Nonuniform in detecting nonuniform DIF was not as high as it was under the condition of equal group abilities.

Table 57: The Type I Error and Power of the LDFA-Nonuniform Procedure in Detecting Nonuniform DIF for the Condition of Unequal Group Ability Distributions (R-N(0,1), F-N(-.5,1))

			Type I Error				**Power**		
20-item Test			Item Discrimination				a-Parameter Difference		
Sample Size			a=0.5	a=1.0	a=1.5		a_D=0.5	a_D=1.0	a_D=1.5
1:1 Ratio	300/300		0.02	0.05	0.07		0.18	0.28	0.38
	600/600		0.06	0.13	0.05		0.31	0.63	0.67
	1200/1200		0.08	0.07	0.08		0.58	0.80	0.86
1:2 Ratio	200/400		0.00	0.12	0.03		0.19	0.25	0.37
	400/800		0.07	0.12	0.06		0.27	0.43	0.66
	800/1600		0.05	0.03	0.03		0.59	0.87	0.84
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.05	0.08	0.05		0.24	0.48	0.53
	600/600		0.07	0.07	0.10		0.45	0.83	0.84
	1200/1200		0.09	0.08	0.07		0.77	0.96	0.99
1:2 Ratio	200/400		0.13	0.05	0.08		0.22	0.41	0.56
	400/800		0.04	0.08	0.06		0.22	0.47	0.62
	800/1600		0.08	0.10	0.13		0.31	0.58	0.83

Unequal Group Ability Distributions **(R-N(0,1), F-N(-1,1))**

The power of LDFA-Nonuniform in detecting nonuniform DIF for the condition of small group ability difference is presented in Table 58 along with its Type I error rates.

No logit analysis was conducted for the case of large group ability difference as the Type I error of LDFA-Nonuniform was high in most conditions and the power seemed to be a result of Type I error rates.

As group ability difference increased, the power of LDFA-Nonuniform decreased. The effect of *a*-parameter difference seemed to be confounded by the effect of group ability difference. This observation is different from those of other procedures, where the power always increased as group ability difference increased. However, for LDFA-Nonuniform, Type I error was also better controlled than for most other procedures.

Table 58: The Type I Error and Power of the LDFA-Nonuniform Procedure in Detecting Nonuniform DIF for the Condition of Unequal Group Ability Distributions (R-$N(0,1)$, F-$N(-1,1)$)

			Type I Error				**Power**		
20-item Test			Item Discrimination				*a*-Parameter Difference		
Sample Size			a=0.5	a=1.0	a=1.5		a_D=0.5	a_D=1.0	a_D=1.5
1:1 Ratio	300/300		0.09	0.10	0.07		0.09	0.18	0.17
	600/600		0.10	0.13	0.12		0.15	0.27	0.38
	1200/1200		0.16	0.25	0.21		0.25	0.51	0.66
1:2 Ratio	200/400		0.10	0.06	0.13		0.07	0.11	0.19
	400/800		0.12	0.17	0.09		0.10	0.23	0.29
	800/1600		0.16	0.28	0.29		0.19	0.34	0.61
40-item Test									
Sample Size									
1:1 Ratio	300/300		0.10	0.13	0.12		0.23	0.37	0.53
	600/600		0.21	0.23	0.13		0.24	0.51	0.62
	1200/1200		0.27	0.39	0.37		0.29	0.51	0.77
1:2 Ratio	200/400		0.09	0.13	0.12		0.10	0.19	0.26
	400/800		0.17	0.17	0.16		0.13	0.43	0.38
	800/1600		0.26	0.36	0.26		0.22	0.52	0.69

In summary, LDFA-Nonuniform had very good control over Type I error when groups were equal in ability and it had inflated but acceptable Type I error in most conditions when

there was a small group ability difference. Under these conditions, LDFA-Nonuniform could be used to detect nonuniform DIF with very high power when group abilities were equal, sample size was medium to large and discrimination was medium to high. However, its power was low when sample size was small or when item discrimination was low. As group ability difference increased, its power decreased. LDFA-Nonuniform was also able to detect balanced uniform DIF to some degree but its power was low in most conditions and it had adequate power only when item discrimination was high and sample size was large.

CHAPTER V
DISCUSSION

The results presented in the previous chapter pertaining to the performance of the Mantel, GMH, STND, SIBTEST, LDFA-Uniform, and LDFA-Nonuniform procedures are discussed below. It is evident from this study that group ability distribution difference was a major factor affecting the performance of each procedure. The results are discussed for the condition of equal group abilities and unequal group abilities separately. Finally, a discussion of the effects of the manipulated factors, i.e., group ability distribution difference, test length, item discrimination, sample size, and sample size ratio is presented.

Equal Group Ability Distributions

Type I Error

When the focal and reference groups were equal in ability, all procedures worked very well in terms of Type I error. The empirical Type I error rates of all procedures were, in general, within or close to the nominal value of .05. The Type I error performances of the Mantel, GMH, STND, LDFA-Uniform and LDFA-Nonuniform procedures were almost indistinguishable. STND performed relatively the best especially when the sample size ratio was 1:2 (with rates ranging from 0 to .02). SIBTEST performed relatively worst, especially with the short test (with rates ranging from .05 to .09), and its Type I error rates were somewhat inflated and slightly higher than the other procedures; however, they were not unacceptable. The mean Type I error rates for the

condition of equal group abilities were .05, .046, .033, .06, .049, and .047 for the Mantel, GMH, STND, SIBTEST, LDFA-Uniform, and LDFA-Nonuniform procedures respectively. Only the mean Type I error rate of SIBTEST was slightly higher than the nominal Type I error rate. The results of logit analyses indicated that test length, item discrimination, and sample size, and sample size ratio (except STND) did not have any significant effect on the Type I error performance of any of the procedures investigated in this study. All procedures performed just as well whether sample size was large or small, test length was short or long, and item discrimination was low or high. It was actually hoped that the independent variables would not affect the rejection rates for items that have no DIF. Only STND was significantly affected by the sample size ratio between the focal and reference groups. Its Type I error rates were substantially lower when the ratio was 1:2 (mean rate of .012) than when the ratio was 1:1 (mean rate of .054). The results of this study in terms of Type I error for equal group abilities were consistent with those of Chang et al. (1996), Spray and Miller (1994), and Zwick, et al. (1993). The procedures they used also performed well when no group ability difference existed.

Since the Type I error was well under control and comparable for all procedures when focal and reference groups were equal in ability, the power of each procedure in detecting DIF in polytomous items can be best assessed and compared under this condition. The comparison of rejection rates of these procedures can be legitimately called a power comparison. As a result, one can have confidence if the observed power of a procedure is high. The performance of each procedure in terms of power in detecting

DIF in polytomous items depends on the type of DIF simulated. In the following, the power of each procedure is discussed for each DIF condition separately.

Uniform DIF (Constant)

The results of this study indicated that Mantel, GMH, STND, SIBTEST, and LDFA-Uniform were extremely powerful in detecting constant uniform DIF. The power of these procedures was almost indistinguishable. This finding is similar to those of Zwick, et al (1993) except they did not use LDFA-Uniform. It was only with the combination of low discrimination and small sample size that these procedures had a lack of power in detecting constant uniform DIF. The LDFA-Uniform was relatively the most powerful procedure, followed by Mantel, SIBTEST, STND, and GMH with GMH being relatively the least powerful. The results of logit analysis indicated that Mantel, GMH, STND, SIBTEST, and LDFA-Uniform were significantly affected by discrimination by sample size interaction as well as the main effects of sample size, sample size ratio (except LDFA-Uniform), and item discrimination. The interaction effect of discrimination by sample size was similar for all procedures. As sample size increased, the difference across the levels of discrimination decreased. Similarly, as discrimination increased, the difference across sample sizes decreased. This interaction is basically due to the ceiling effect observed as item discrimination and sample size increased. As for the main effects, in general, the power increased as sample size and item discrimination increased and the power decreased slightly when the sample size ratio between the focal and reference groups changed from 1:1 to 1:2 (except LDFA-Uniform). The power of STND decreased considerably as the ratio changed to 1:2. The sample sizes of 300/300

and 200/400 appeared to be inadequate especially when item discrimination was low. For the GMH and STND procedures, the power was even inadequate when item discrimination was medium. The effects of item discrimination and sample size are as expected. It is a well known fact that the statistical power is directly related to sample size, with greater power being associated with a larger sample size. Also for a fixed shift in item difficulty, the amount of DIF in the item is directly dependent on the discrimination parameter.

For STND, both the Type I error and power seemed to be more affected by sample size ratio between the focal and reference groups than the other procedures. Comparing samples of the same total number of examinees, higher Type I error and power were associated with the ratio of 1:1 than with the ratio of 1:2. Although significant statistically, this effect of sample size ratio was really not that important when two groups had equal abilities because both ratios produced Type I error rates within or close to the nominal level. The lower Type I error and power observed at the 1:2 ratio is probably because in calculating the STND statistic, focal group weighting was used, i.e., the between-group difference in mean item score for examinees at each level of the matching variable was weighted by the proportion of focal group members who were at that level of the matching variable. Another procedure in which a weighting function is used is SIBTEST. In this study, in computing SIBTEST statistic, weights based on the total group were used. Shealy and Stout (1993) reported that total group weighting produced better Type I error control in their simulation studies, which seemed not to be the case with this study.

As expected, the LDFA-Nonuniform procedure was unable to detect constant uniform DIF since no interactions of either group by ability or group by category were simulated into the studied items in the uniform DIF (constant) condition. The results of LDFA-Nonuniform were comparable to its Type I error rates.

Uniform DIF (Balanced)

In this condition, the *a*-parameter for each group remained the same and the *b*-parameters varied, but in different directions, resulting in nonunidirectional DIF. What was simulated was uniform DIF within the response category that was not constant across the response categories, i.e., a group by response category interaction.

The power of Mantel, STND, SIBTEST, and LDFA-Uniform in detecting balanced uniform DIF was a reflection of their Type I error pattern. They failed to detect balanced uniform DIF almost completely when two groups had equal abilities. This is not surprising since they are not designed to detect interactions of any kind. They are sensitive to mean differences, which are anticipated to be minimal in this type of DIF (because of canceling out effect). They are all designed to detect unidirectional DIF, whereas balanced uniform DIF changes sign across score categories. This finding is consistent with those of Chang et al. (1996) and Zwick, et al. (1993). They found that the Mantel, STND, and SIBTEST were unable to detect uniform DIF of balanced type. Spray and Miller (1994) also found that the Mantel (called MH_{ord} in their paper) and LDFA-Uniform procedures failed completely to detect balanced uniform DIF (labeled as nonuniform DIF in their paper) even with sample size as large as 2000 when the focal and reference groups had identical ability distribution.

Both the GMH and LDFA-Nonuniform procedures were able to detect balanced uniform DIF to some degree with GMH being much more powerful. When two groups were equal in ability, GMH demonstrated a high degree of power when sample size was large and item discrimination was medium to high or when discrimination was high and sample size was medium. Its power was inadequate in other conditions and was even very low when item discrimination was low or when sample size was small. The power of LDFA-Nonuniform was low in most conditions except with the combination of large sample size and high discrimination. The results of logit analyses indicated that both procedures were significantly affected by item discrimination and sample size. In general, their detection rates increased as sample size and item discrimination increased. In addition, GMH was also affected by the sample size by item discrimination interaction, which was basically due to the ceiling effect observed at larger sample size and higher discrimination. The findings of this study are consistent with those of Spray and Miller (1994) and Zwick, et al. (1993). Spray and Miller (1994) found that when two groups had identical ability distributions, both GMH (called MH_{nom} in their paper) and the LDFA-Nonuniform procedures detected balanced uniform DIF with 100% accuracy at sample size of 2,000. In their study, the GMH procedure showed very high power even with the smaller sample size (500); the LDFA-Nonuniform DIF test had a low-to-moderate degree of power at the same sample size. Zwick, et al. (1993) found that the GMH was much superior to the Mantel in detecting balanced uniform DIF and the GMH is sensitive to between-group differences in the frequencies of any of the item scores instead of between-group mean differences. They recommended that if the entire response

distribution, rather than merely the means, is of interest, the GMH should be used even when the response categories are ordered.

Nonuniform DIF

The Mantel, STND, SIBTEST, and LDFA-Uniform procedures were not useful in detecting items with nonuniform DIF when the reference and focal groups had identical ability distributions. This was expected since these procedures are not designed to detect nonuniform DIF. The detection rates of these procedures were reflections of their Type I error patterns and they increased as group ability difference increased. When group ability differences existed, these procedures produced spurious results.

Both the LDFA-Nonuniform and GMH procedures were able to detect nonuniform DIF with the former being much more powerful than the latter when groups were equal in ability. The power of LDFA-Nonuniform was adequate to high when the *a*-parameter difference was medium to large or when sample size was medium to large but its power was inadequate when *a*-parameter difference was small (except with large sample size) or when sample size was small. The power of GMH was low in most conditions except when the *a*-parameter difference was medium to large and sample size was large. The logit analyses indicated that both procedures were significantly affected by the *a*-parameter difference by sample size interaction as well as the main effects. As usual, the differences among the three levels of *a*-parameter difference decreased as sample size increased. The detection rates of both procedures increased as sample size and *a*-parameter difference increased. In addition, GMH was also affected by test length and its power increased as test length increased. Spray and Miller (1994) found that when

two groups had identical ability distributions, the GMH procedure was much more powerful than the LDFA-Nonuniform procedure in detecting nonuniform DIF at the sample size of 500 but the two procedures were equally powerful at the sample size of 2000. The findings in this study do not coincide with those of Spray and Miller (1994). In this study, when two groups had equal abilities, the LDFA-Nonuniform procedure was much more powerful than the GMH procedure in all conditions.

Unequal Group Ability Distributions

Type I Error

When the focal and reference groups differed in ability, the Type I error rates of all procedures increased considerably. When item discrimination was low, group ability difference seemed not to be a big issue. It was when item discrimination was medium to high that group ability difference became a major issue. The Type I error rates (except LDFA-Nonuniform) of all procedures were within or close to .05 when item discrimination was low regardless of group ability difference. This is probably because items with low discrimination were poor in discriminating between examinees with low and examinees with high abilities. Although the Type I error of these procedures was in good control at low discrimination, the power of these procedures in detecting DIF was lower except with large sample size.

When there was a small group ability difference, the Type I error rates were somewhat inflated but still acceptable (<.10) at medium item discrimination except for large sample size for Mantel, GMH, STND (when the ratio was 1:2), SIBTEST, and

LDFA-Uniform; these procedures had high power in detecting uniform DIF when item discrimination was medium but their power should be interpreted with caution especially when sample size was large. The Type I error of STND for the sample size ratio of 1:1 seemed a little too high especially with the 20-item test. The Type I error rates were highly inflated and unacceptable (>.10) at high discrimination for Mantel, GMH (except with the 40-item test), STND (except with the 40-item test and sample size ratio of 1:2), SIBTEST, and LDFA-Uniform. This caused one to question their power under the same condition. LDFA-Nonuniform performed relatively better than other procedures under the condition of small group ability difference as its Type I error inflation was much less severe than that of other procedures. It may be used when the group ability difference is small but the result should be interpreted with caution . For STND, again, the ratio of 1:2 produced substantially lower Type I error rates than the ratio of 1:1 as its Type I error rates were within or close to .05 in most conditions when the ratio was 1:2; therefore, STND may be appropriate to use when the group ability difference is small using the 1:2 sample size ratio.

When there was a large group ability difference, the Type I error of all procedures, except LDFA-Nonuniform, was only acceptable (<.10) at low item discrimination and was highly inflated when item discrimination was medium and was out of control when item discrimination was high. Consequently, one should not put any faith in the power of these procedures when item discrimination was medium to high. In other words, these procedures are not appropriate to use when groups differ substantially and item discrimination was high. For LDFA-Nonuniform, the power was questionable for all

levels of item discrimination as its Type I error was unacceptable when a large group ability difference was present.

Contrary to what was found for equal group abilities , the Type I error rates of SIBTEST appeared to be relatively lower than those of the other procedures when the group ability distribution differed, especially when item discrimination was high. It might be useful when differences in the reference and focal groups ability distributions exist in practical settings. This is consistent with the findings of Zwick et al. (1997); they found that the Type I error of SIBTEST was higher than Mantel and STND when equal group ability distribution was used and the Type I error of SIBTEST was lower than Mantel and STND when unequal group ability distribution was used. This appears to be due to the regression correction used in the SIBTEST procedure. Comparing Mantel and GMH, it was found that GMH had relatively better control over Type I error when two groups were different in terms of their ability distributions. This is probably because Mantel compares groups with respect to their means but GMH compares groups with respect to their entire item response distributions rather than just item means. As a results, Mantel could be more sensitive to group ability difference.

In general, all procedures had good or acceptable Type I error control when groups were equal in ability or when there was a small group ability difference but item discrimination was not high or when there was a large group ability difference but item discrimination was low (except LDFA-Nonuniform which had acceptable Type I error when there was a small group ability difference but totally unacceptable Type I error when there was a large group ability difference). In addition, STND seemed to work well

when there was a small group ability difference and sample size ratio was 1:2. These procedures could be used with confidence under the above conditions. However, the results should be interpreted with caution when there is a small group ability difference, sample size is large, and item discrimination is medium. These procedures are not appropriate to use with the combination of small group ability difference and high discrimination (for GMH, with the short test length; for STND, with sample size ratio of 1:1 and short test length) and the combination of large group ability difference and medium to high discrimination due to the excessive Type I error under these conditions. In the following, the comparison of the power of each procedure is made with respect to the conditions discussed above and under which their Type I error was in good control or at least acceptable.

Uniform DIF (Constant)

Consistent with the results of the Type I error study, the power of Mantel, GMH, STND, SIBTEST, and LDFA-Uniform in detecting constant uniform DIF also increased as group ability difference increased and was very high in most conditions. However, their power was inadequate with the combination of low discrimination and small sample size. In addition, STND (sample size ratio of 1:2) and GMH also had a lack of power with the combination of low discrimination and medium sample size. As discussed in the previous section, these procedures had highly inflated Type I error with the combination of small group ability difference and high discrimination and the combination of large group ability difference and medium to high discrimination, therefore, the power under these conditions was questionable. The Mantel, GMH, STND, SIBTEST, and LDFA-

Uniform procedures can be used to detect constant uniform DIF with confidence when groups are equal in ability, or when groups differ in ability but item discrimination is low; they may also be used when there is a small group ability difference and item discrimination is medium. The results should be interpreted with caution as they are inappropriate to use with the combination of small group ability difference and high discrimination (for GMH, unless test length is long; for STND, unless sample size ratio is 1:2 and test length is long) and the combination of large group ability difference and medium to high discrimination.

Uniform DIF (Balanced)

The power of Mantel, STND, SIBTEST, and LDFA-Uniform in detecting balanced uniform DIF was a reflection of their Type I error pattern and both increased as group ability difference increased. When group ability differences existed, these procedures could produce misleading results.

Both GMH and LDFA-Nonuniform were able to detect balanced uniform DIF to some degree when there was no group ability difference. As group ability difference increased, their power increased substantially with the GMH being more powerful in all conditions. This indicated that GMH was more sensitive to directional change across the response categories. The Type I error of GMH was unacceptable with the combination of small group ability difference, high discrimination, and short test length, and the combination of large group ability difference and medium to high discrimination. Therefore, GMH is not appropriate to use under these conditions. It may be used when item discrimination is low to medium and group ability difference is small or when item

discrimination is low and group ability difference is large. Its power was very low when item discrimination was low or when sample size was small.

LDFA-Nonuniform may be used to detect balanced uniform DIF when group ability difference is small but the results should be interpreted in conjunction with its Type I error rates which were somewhat inflated. Its power was very low except when item discrimination was high and sample size was medium to large. It should not be used when groups differ substantially in ability as its Type I error was unacceptable under this condition.

Nonuniform DIF

The power of Mantel, STND, SIBTEST, and LDFA-Uniform in detecting nonuniform DIF reflected their Type I error pattern which increased as group ability difference increased. When group ability differences existed, these procedures could produce misleading results.

Both LDFA-Nonuniform and GMH were able to detect nonuniform DIF, but group ability distribution difference had a different impact on the power of these two procedures. When two groups were equal in ability, LDFA-Nonuniform was much more powerful than GMH. The power of LDFA-Nonuniform decreased substantially as the between-group ability differences increased. It could be that the influence of nonuniform DIF was confounded or obscured by group ability difference. However, its Type I error was still acceptable when a small group ability difference was present. Nevertheless, its Type I error was totally unacceptable when a large group ability difference was present. On the contrary, the power of GMH increased substantially as the between-group ability

differences increased as did its Type I error. When the focal group ability was substantially lower than the reference group ability, GMH had much higher detection rates than LDFA-Nonuniform.

LDFA-Nonuniform may be used to detect nonuniform DIF when group ability difference is small but the results should be interpreted in conjunction with its Type I error rates which were somewhat inflated. Large or at least medium sample size should be used to provide adequate power; it should not be used when groups differ substantially in ability as its Type I error was unacceptable under this condition.

The Type I error of GMH was unacceptable with the combination of small group ability difference, high discrimination, and 20-item test, and the combination of large group ability difference and medium to high discrimination. Therefore, GMH is not appropriate to use under these conditions. It may be used when item discrimination is low to medium and group ability difference is small or when item discrimination is low and group ability difference is large. Its power was very low under these conditions and large sample sizes should be used.

The Effect of the Manipulated Factors

Group Ability Distribution Difference

The effect of group ability distribution difference on the performance of the Mantel, GMH, STND, SIBTEST, LDFA-Uniform, and LDFA-Nonuniform has already been discussed in the previous sections. All procedures worked well when there was no group ability difference but had Type I error inflation when group ability difference

existed. The larger the group ability difference, the more severe the Type I error inflation. In general, the power as well as Type I error increased as group ability difference increased.

Test Length

In this study, the Type I error and power of all procedures were not significantly affected by test length. This is probably because the two test lengths used in this study are both long enough to produce reliable test scores. However, the Type I error rates were slightly lower for the 40-item test than for the 20-item test and power was higher for the 40-item test than for the 20-item test. This is consistent with the theory that longer tests tend to be more reliable than the shorter tests and any distortions in the functioning of the statistical procedures are expected to be less severe with a longer test (when reliability is higher) than with a shorter test. LDFA-Nonuniform was less influenced by test length than other procedures. Test length had a slight effect on its Type I error rates only when there was a substantial group ability difference. LDFA-Nonuniform appeared to be affected by test length in a different way from other procedures and both its Type I error and power increased as test length increased.

Sample Size

In this study, increasing sample size generally led to an increase in Type I error rates and power for all procedures. This effect of sample size was expected. When sample size was large, even small effects could be detected and thus the Type I error of a statistic may be inflated by large sample sizes and higher power of a statistic at large sample size could be due to high Type I error rates. In this study, it was found that when there existed

a group ability difference, sample sizes of 1200/1200 and 800/1600 produced much higher Type I error rates than other sample sizes for all the procedures under investigation. Sample sizes of 300/300 and 200/400 (except when they were combined with high discrimination) appeared to be to small for Mantel, GMH, STND, SIBTEST, and LDFA-Uniform to have adequate power in detecting constant uniform DIF and for GMH and LDFA-Nonuniform to have adequate power in detecting balanced uniform DIF or nonuniform DIF. It seemed that sample sizes of 600/600 and 400/800 achieved a balance between reasonable Type I error rates and adequate power. However, if the studied-item discrimination is low, large sample size is recommended.

Sample Size Ratio

The performance of STND was influenced by sample size ratio between the focal and reference groups. Comparing samples of the same total number of examinees, higher Type I error and power were associated with the ratio of 1:1. As has already been explained, this is probably because that focal group weighting was used in calculating the STND statistic. It seems that both the 1:1 and 1:2 ratios can be used as both produced Type I error within or close to the nominal value of .05; however, when a small group ability difference was present, sample size ratio of 1:2 is preferred as it provides better control over Type I error although with a loss of power.

Item Discrimination

Except for LDFA-Nonuniform, all other procedures were significantly affected by discrimination parameters of the studied items. In general, the Type I error rates and power increased as item discrimination increased. This is consistent with the theory that

effective amount of DIF of an item depends on both its difficulty and discrimination parameters. When item discrimination was low, the Type I error rates of all procedures except LDFA-Nonuniform were within or close to .05 regardless of the levels of ability distribution difference, test length, sample size, and sample size ratio. This is probably because that an item with low discrimination was not as capable of distinguishing between low- and high-ability examinees. LDFA-Nonuniform had high Type I error even when item discrimination was low. As item discrimination increased, the Type I error rates were inflated and became out of control especially when there was a substantial group ability difference. In addition, as item discrimination increased, the Type I error rates of these procedures were more inclined to be affected by ability distribution difference. Similarly, the power of all procedures was low when item discrimination was low and their power increased as item discrimination increased. The findings regarding the effect of item discrimination on Type I error of DIF detection procedures were not consistent with those of Chang et al. (1996). They found that the Type I error rates of Mantel and STND were lower when item discrimination was in the medium range and were higher when item discrimination was either extremely low or extremely high. In this study, although less dramatically than other procedures, SIBTEST was affected by item discrimination to a fairly high degree especially when the group ability distribution difference was large. Nevertheless, Chang et al. (1996) found that SIBTEST was considerably more robust against item discrimination than Mantel and STND and its Type I error did not increase or decrease as dramatically as that of Mantel and STND.

CHAPTER VI
CONCLUSIONS

The conclusions pertaining to this investigation are presented in this section. This chapter is divided into three sections. In the first section, an overall conclusion pertaining to the performance of the Mantel, GMH, STND, SIBTEST, and LDFA procedures in detecting DIF in polytomous item responses is presented. In the second section, some limitations of this study are identified. Finally, some recommendations for future research are provided in the last section.

Overall Conclusion

The performance of the Mantel, GMH, STND, SIBTEST, and LDFA (LDFA-Uniform and LDFA-Nonuniform) procedures was found to be affected by group ability distribution difference. Both the Type I error and power of these procedures increased as the between-group ability difference increased, except for LDFA-Nonuniform for which the power in detecting nonuniform DIF decreased as group ability difference increased.

When two groups were equal in ability, the Mantel, GMH, STND, SIBTEST, and LDFA procedures had good control over Type I error with STND performing relatively the best and SIBTEST performing relatively the worst. Except for STND, the Type I error of all the other procedures investigated in this study was not affected by sample size, sample size ratio, test length, or item discrimination. STND was affected by sample size ratio. Smaller than the reference group sample size is preferable for the focal group for

STND because it produced lower Type I error rates, however with the sacrifice of low power.

When two groups had a small ability difference, all procedures generally had good control over Type I error when item discrimination was low, inflated but still acceptable Type I error when item discrimination was medium, but had highly inflated and unacceptable Type I error when item discrimination was high (except LDFA-Nonuniform).

When two groups had a large group ability difference, except for LDFA-Nonuniform, when item discrimination was low, all the other procedures had good control over Type I error and were comparable; however, when item discrimination was medium to high, the Type I error of all procedures was generally unacceptable. LDFA-Nonuniform had unacceptable Type I error for all levels of item discrimination.

Based on the findings of this study, the Type I error of all procedures was generally unacceptable with the combination of small group ability difference and high discrimination and the combination of large group ability difference and medium to high discrimination. One should question the power of the procedures under these conditions. They are not appropriate to use under these conditions.

Based on the findings of this study, all procedures had good and comparable Type I error control when groups were equal in ability or when groups had a small ability difference and item discrimination was not high or when groups had a large ability difference and item discrimination was low; therefore, the following conclusions regarding the power are made based on the above conditions.

If the purpose of the DIF study is to only identify constant uniform DIF, all procedures (except LDFA-Nonuniform) investigated in this study can be used. They were all powerful in detecting constant uniform DIF, with the LDFA-Uniform being the most powerful, followed by the Mantel, SIBTEST, STND, and GMH. However, if the studied-item discrimination is low, at least medium sample size should be used.

If any group-by-response-category interaction (i.e., balanced uniform DIF) is suspected, The GMH and LDFA-Uniform are the preferred procedures with GMH being much more powerful. However, GMH was not very powerful when sample size was low or when item discrimination was low and LDFA-Nonuniform was not very powerful when item discrimination was low to medium and sample size was small to medium. Therefore, large sample size should be used for LDFA-Nonuniform and at least medium sample size should be used for GMH when detecting balanced uniform DIF. The results should be interpreted with caution because of the possibility of high Type I error.

When nonuniform DIF is of interest, the LDFA-Nonuniform and GMH procedures should be used. They were able to detect nonuniform DIF with LDFA-Nonuniform being more powerful when two groups had equal ability distributions and GMH being more powerful when two groups had different ability distributions. However, LDFA-Nonuniform was not very powerful when sample size was low or when item discrimination was low and GMH was not very powerful when item discrimination was low to medium and sample size was small to medium. Therefore, large sample sizes (e.g. 1200/1200) should be used for GMH and at least medium sample sizes (e.g. 600/600)

should be used for LDFA-Nonuniform, but the results should be interpreted with caution due to the possibility of high Type I error.

Regarding sample size, it appeared that sample sizes of 1200/1200 and 800/1600 produced much too high Type I error rates when there was a group ability distribution difference, and sample sizes of 300/300 and 200/400 were inadequate for the power of DIF detection procedures especially when item discrimination was low. Sample sizes of 600/600 and 400/800 seemed to produce a balance between reasonable Type I error and adequate power and therefore were recommended. With respect to sample size ratio between the focal and reference groups, both ratios used in this study worked well. However, for STND, the ratio of 1:2 produced relatively lower Type I error than the ratio of 1:1, but the power was also lower for the ratio of 1:2. As for test length, both two test lengths seemed to be appropriate based on the results of this study. Generally, the longer test produced lower Type I error and higher power for the DIF detection procedures. In terms of studied-item discrimination, the Type I errors of the DIF detection procedures in this study all seemed to be within the theoretical nominal value when item discrimination was low but were seriously inflated at high item discrimination when there was a group ability difference. The results of a DIF study should be interpreted with caution when item discrimination is high as there is a possibility of high Type I error.

Finally, results from this study indicated that the LDFA procedure is preferable over the other procedures. It had good adherence to Type I error when two groups had equal ability distributions. Most important of all, it can distinguish between uniform and nonuniform DIF and can detect both types of DIF. The fact that it can also detect

balanced uniform DIF indicates that it may be capable of detecting interactions of any kind, whether group by ability or group by response category. However, GMH might be preferable for balanced uniform DIF.

Limitations of the Study

Some of the limitations within which this study was conducted can be summarized as follows:

1. The advantages of a simulation study are that the Type I error and power of statistical procedures can be evaluated based on known parameters. However, conducting a simulation study limits the extent of generalization of the results to the conditions that were examined in this study.

2. This study was limited to the analysis of DIF that was introduced in only one item of the test.

3. In this study, the difference in sample size between the focal and reference groups was set at one-to-one and one-to-two ratios. It is acknowledged that in many testing situations, group numbers may be more disparate.

4. This study did not incorporate a systematic investigation of the effects of jointly varying the discrimination and difficulty parameters of the studied items for the nonuniform DIF. In the future, it would be interesting to examine this effect.

5. In this study, only one DIF magnitude of 0.25, was simulated in the studied items. In practice, larger or smaller DIF magnitude than the one used in this study may be found in real testing situations.

6. The number of replications, in this study, for each condition was set at 100 replications. Some researchers might consider this to be too small to yield precise results. From the results of Type I error study, it seemed that more than 100 replications would eliminate certain ambiguities. However, 432 conditions were used in this study for each of the six (two for LDFA) DIF detection procedures. Based on practical considerations, 100 replications were considered to be adequate to get a general idea of the performance of the DIF detection procedures.

Suggestions for Future Research

In light of the findings of this study, the following suggestions are offered regarding future investigations that might be undertaken in this area of research.

1. It would seem to be important to assess the effectiveness of the DIF detection procedures under conditions other than those used in this study, such as different group ability distribution, test length, sample size, sample size ratio between focal and reference groups, item discrimination, item difficulty, the amount of DIF, etc. More work is needed to understand exactly what conditions tend to produce the best and worst performance for the DIF detection methods and to develop guidelines under which these methods might be used.

2. The accuracy of the DIF detection procedures should be investigated when matching tests are contaminated with DIF items. This is expected to affect the accuracy of the ability estimates and, therefore, the effect of test length on the accuracy of the DIF detection procedures may be more important.

3. In the future, it would also be interesting to investigate the effect of having more than one DIF item in a test or to analyze a group of DIF items for simultaneous DIF.

4. It would be interesting to investigate the effects of jointly varying the discrimination and difficulty parameters of the studied items.

5. In the future study, other plausible DIF magnitudes and DIF types should be assessed.

6. It must be remembered that DIF analyses, while they can be helpful in investigating the accuracy of statistical DIF detection procedures, are only one component of the extensive research that is needed on the validity and fairness of tests.

7. Finally, it should be pointed out that conclusions regarding the performance of the DIF detection procedures investigated in this study were limited to the conditions examined in this study. Generalizability to actual test data or simulated data based on conditions other than those examined in this study may provide different results.

References

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

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Allen, N. L., & Donoghue, J. R. (1994). *DIF analysis based on complex samples of dichotomous and polytomous items*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Allen, N. L., & Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement, 33*, 231-251.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Baker, F. B. (1981). Log-linear, logit-linear models: A didactic. *Journal of Educational Statistics, 6*, 75-102.

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias*. Baltimore: Johns Hopkins University Press.

Bock, R. D. (1975). *Multivariate Statistical Methods in Behaviroal Reserch*. New York, NY: McGraw-Hill.

Bolt, D., & Stout, W. (1995). *Analysis of DIF of dichotomously and polytomously scored items*. Invited survey paper to appear in Behaviormetrika.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks: Sage Publications, Inc.

Chang, H.-H., & Mazzeo, J. (1994). The unique correspondence of the item response function and the item category response functions in polytomously scored item response models. *Psychometrika, 59*, 391-404.

Chang, H.-H., Mazzeo, J., & Roussos, L. A. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333-353.

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

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

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dorans, N. J., & Kulick, E. M. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic perspective. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135-165). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Douglas, J., Stout, W., & Dibello, L. (1996). A kernel smoothed version of SIBTEST with applications to local DIF inference and function estimation. *Journal of Educational and Behavioral Statistics, 21*, 333-363.

Gong, Y (1998). *A modification to PSYCH.FOR and GENMH.FOR (Spray, 1994).*

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315-332.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, Sage Publications.

Hanson, B. A. (1992). *Comments on Miller and Spray's paper*. Internal Memorandum ACT Program.

Hanson, B. A., & Feinstein, Z. S. (1995). *A polynomial loglinear model for assessing differential item functioning*. ED384629.

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.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Johnson, E. G., & Carlson, J. E. (1994). *The NAEP 1992 technical report.* Washington, DC: National Center for Education Statistics.

Lewis, C. (1993). A note on the value of including the studied item in the test score when analyzing test items for DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.317-319). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58,* 690-700.

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

Masse, L. C. (1994). *A presentation and comparison of some new statistical techniques in the analysis of polytomous differential item functioning: A Monte Carlo investigation.* Unpublished Ph.D. thesis, University of Ottawa, Ottawa, Canada.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1991, April). *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.

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1994, April). *Multidimensional DIF analyses: The effects of matching on two internally derived ability scores.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Mazzeo, J., & Chang, H.-H. (1994, April). *Detecting DIF for polytomously scored items: Progress in adaptation of Shealy-Stout's SIBTEST procedure*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*, 289-311.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*, 107-122.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16*, 389-402.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-328.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

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

Rogers, H. J., & Swaminathan, H. (1993, April). *Differential item functioning procedures for non-dichotomous responses*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Roussos, L.A., & Stout, W. F. (1993, April). *Simulation studied of effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*, 34 (4, Pt. 2).

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education, 2,* 255-275.

Shealy, R. (1989). *An item response theory-based statistical procedure for detecting concurrent internal bias in ability tests.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separate true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Spray, J., & Miller, T. (1994). *Identifying nonuniform DIF in polytomously scored test items.* ACT Research Report Series 94-1.

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

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 147-170). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18,* 15-25.

Welch, C., & Hoover, H. D. (1993). Procedures for extending item bias techniques to polytomously scored items. *Applied Measurement in Education, 6,* 1-19.

Welch, C., & Miller, T. R. (1995). Assessing differential item functioning in direct writing assessment: problems and an example. *Journal of Educational Measurement, 32,* 163-78.

Wilson, A. W., Spray, J. A., & Miller, T. R. (1993, April). *Logistic regression and its use in detecting nonuniform differential item functioning.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Zwick, R., Donoghue, J. r., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21*, 187-201.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*, 321-344.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18*, 121-140.