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.

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

Université d'Ottawa • University of Ottawa

Detecting DIF in Polytomous Items: An Empirical Comparison of the Ordinal Logistic Regression, Logistic Discriminant Function Analysis, Mantel, and Generalized Mantel Haenszel Procedures

by
Elizabeth Kristjansson
Faculty of Education, University of Ottawa

Submitted in partial fulfilment of the
requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies and Research
University of Ottawa

ACKNOWLEDGMENTS

Completing this dissertation has been a lesson in endurance and perseverance. I could not have possibly achieved this without the support of my family, mentors, and friends. I would like to thank those who made it possible for me to achieve this goal, and who helped me to get through the past six years.

Special thanks to:

- My husband, Gary, and older sons, Erik and Matt, for much support and understanding and for their assistance with housework and babysitting. Thanks to Kris for being a wonderful kid.

- The ladies of the CSHA: E.Hay, Andrea, Maggie, Liz, Joan, and Linda, for many good years of friendship and laughter.

- My mentors in the CSHA from across the country, particularly Holly, Gerry, John, Parminder, and Larry. Your advice and encouragement have helped me to push on.

- My parents and grandparents for giving me unconditional love, and for instilling a lifelong love of learning in me.

- My father-in-law and mother-in-law for their unwavering support and pride in me.

- Brad Cousins and Marielle Simon for your encouragement, support, and insightful comments and suggestions on this dissertation.

- Marvin Boss for sharing his time, wisdom, and expertise with me. I enjoy working with you, and have learned a great deal from you. I'm grateful that I got the chance to know you and your wonderful wife.

- Richard Aylesworth, master computer programmer and true friend. Your expertise, dedication, and encouragement helped to make this possible.

- Finally, to Ian McDowell, teacher, mentor, and friend. Throughout the past twelve years, you have taught me well, stimulated my intellect and expanded my mind, and encouraged me to push myself further than I thought that I can go. Thank you; I hope that we can work and learn together for many more years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Bias is an important threat to item and test validity. Differential item functioning (DIF) refers to the statistical procedure for the identification of potentially biased items. The assessment of DIF is an essential step in the validation of educational tests. Techniques for assessing DIF in dichotomously scored items are well established. However, with educational reform, there has been a move away from the tradition of tests comprised solely of binary items toward performance-based assessment, which includes polytomous items. DIF in polytomous items is more complicated than DIF in dichotomous items, and techniques for identifying DIF in polytomous items have not yet been subjected to thorough empirical study. Most of the existing polytomous DIF detection procedures have been evaluated in only a few studies. Prior to this study, Ordinal Logistic Regression had not been empirically evaluated at all. There is a need to assess the performance of polytomous DIF detection procedures under a variety of different conditions which may occur in applied situations. This forms the goal of this thesis.

This study had four main objectives and two minor objectives: 1) To compare the Type I Error rates of four analytic techniques: the Mantel, Generalized Mantel-Haenszel (GMH), Logistic Discriminant Function Analysis (LDFA), and Ordinal Logistic Regression (OLR) procedures when there was no DIF in items. It was hypothesized that the procedures would differ little in their Type I Error rates, but that the OLR would have relatively the lowest Type I Error rates. 2) To compare the power of the Mantel, GMH, LDFA, and OLR for detecting uniform and nonuniform DIF in polytomous items. It was hypothesized that the Mantel would have the highest power for uniform DIF, but that it would not be useful for detecting nonuniform DIF. The GMH,

OLR and LDFA were expected to display high power for nonuniform DIF 3) To learn whether discrimination of the studied item, reference and focal group ability difference, sample size ratio between reference and focal group, and skewness would affect the performance of the four DIF detection procedures. It was hypothesized that differences in group ability distributions would result in increased Type I Error when item discrimination was high, and that differences in sample size ratio would lower power. 4) To determine whether adding a measure of effect size would reduce Type I Error. It was hypothesized that including effect size in the decision rule would reduce Type I Error for all procedures, and that it would also result in slightly lower power. 5) To compare Type I Error of the LDFA and OLR in classifying DIF as uniform when it was nonuniform and 6) To compare the Type I Error of the LDFA and OLR in classifying DIF as nonuniform when it was uniform.

In order to meet these objectives, a 26- item test was simulated. The first 25 items were always free of DIF; the 26th item was the studied item, and it had no DIF, uniform DIF, or nonuniform DIF. To simulate uniform DIF, the difference in *b*-parameters between the focal and reference groups at each score threshold was set at a constant 0.25. To simulate nonuniform DIF, the difference in *a*-parameters between the reference and focal groups was set at 1.0, 1.3, and 1.6. Several variables which might affect performance of the four techniques were studied; these included the discrimination of the studied item, the ability level of the focal and reference groups, sample size ratio, and the skewness of the ability distributions. Three levels of studied item discrimination were assessed: low (0.8), moderate (1.2), or high (1.6). Two differences in ability levels of the reference and focal group were studied: no difference, and a small ability difference (focal group ability was 0.5 standard deviations lower than reference group ability). The ratio of

the sample sizes in the reference and focal groups was set to be either equal (2000:2000) or 4:1 (3200:800). Finally, the ability distributions of the reference and focal groups were either normal, or were moderately (- 0.75) negatively skewed. These permutations yielded 72 study conditions under which to compare the four analytic methods. 400 replications were done for each condition.

Two new measures of effect size (one for uniform DIF and one for total DIF) were developed by Aylesworth and Kristjansson (2000) in conjunction with the OLR; both were used in the decision rules for OLR. The effect size for total DIF was also used in the decision rules for the GMH, and the effect size for uniform DIF was used in the decision rules for the Mantel. With these decision rules, the item had to show both statistical significance and practical significance (predicted group difference in item scores had to exceed 0.3 points out of a 3-point scale). Similarly, measures of effect size for uniform DIF and for total DIF were also developed for LDFA.

The Type I Error and power for uniform and nonuniform DIF was assessed under all study conditions in order to compare the four procedures. OLR and LDFA were also compared on their ability to discriminate between uniform and nonuniform DIF. To study the impact of effect size, the difference in Type I Error and power of each procedure with and without effect size was calculated. Logistic regression was used to study the effect of the independent variables (sample size ratio, item discrimination, group ability differences, and skewness) on the Type I Error and power for all four procedures.

Several interesting findings emerged from this study. As hypothesized, the four procedures were almost indistinguishable in terms of their Type I Error; the mean of each one was below the nominal value of 0.05. However, the hypothesis that OLR would have the lowest Type

I Error was not upheld; instead the Mantel had the lowest average Type I Error. The Type I Error of the LDFA was somewhat more problematic than that for the other techniques; its mean Type I Error reached levels that were greater than 0.05 in thirteen of twenty-four conditions. Similarly, each of the four procedures had high and nearly equivalent power for uniform DIF, although the mean power for the LDFA was the highest at 0.99 and the mean power for the OLR was the lowest at 0.96. The hypothesis that the Mantel would have the highest power for uniform DIF was not upheld; although its power was extremely high, the LDFA had a very slightly higher mean power for uniform DIF. The four procedures differed widely in terms of detecting nonuniform DIF. As hypothesized, the Mantel could not detect nonuniform DIF while the other three procedures could do so. However, the LDFA had much lower power for nonuniform DIF than the OLR and GMH. The GMH and OLR had equivalent power for nonuniform DIF under all study conditions.

High item discrimination resulted in slightly higher power for uniform DIF for most procedures. Interestingly, moderate and high item discrimination resulted in substantially lower power for the LDFA for nonuniform DIF. Higher (4:1) sample size ratios resulted in somewhat lower power for all procedures for both uniform and nonuniform DIF. The small differences in group ability distribution simulated in this study had generally little effect, except that there was an interaction between group ability difference and item discrimination so that Type I Error was somewhat higher for the Mantel, GMH, and LDFA procedures for group ability difference and high discrimination. Skewness had little effect on the power or on the Type I Error of these procedures.

As hypothesized, inclusion of an effect size measure into the procedure for DIF detection

slightly reduced the Type I Error of the Mantel, GMH, and OLR for null DIF. However, it resulted in very large improvements in the classification of DIF as uniform or nonuniform.. Inclusion of effect size resulted in slight reductions in power for nonuniform DIF for the GMH and OLR; effect size did not change the power of the LDFA. It was concluded that, overall, the OLR and the GMH were the best procedures for detecting DIF in polytomous items. The OLR has the advantage of being able to discriminate between different types of DIF. Because it cannot detect nonuniform DIF, the Mantel should not be used, except possibly in the second phase of DIF detection, to discriminate between uniform and nonuniform DIF. LDFA may be easier to use than OLR, but should only be used when studied item discrimination is low and there are no differences in group ability.

CHAPTER I
INTRODUCTION

The ability to do well on educational tests increasingly affects the life and material chances of children and adults in our society (Camilli & Shepard, 1994). Standardized testing is now being used as a way of selecting for educational placement or admission to university and job training, for identifying people with special needs, and for evaluating educational programs. Those of us who develop, administer, and interpret tests are morally bound to ensure that the conclusions and decisions made on the basis of testing are valid (Camilli & Shepard, 1994; Holland & Wainer, 1993).

Bias is the major threat to test validity. If an item or whole test is biased, examinees are denied the chance to demonstrate their true abilities and incorrect inferences and decisions may be made on the basis of such tests. Bias may be shown against one gender, or a particular linguistic, racial, or social group. For example, in testing for giftedness, bias could result in members of a particular group being falsely screened out of gifted programs. Bias detection procedures are designed to detect or uncover such differential item validity (Camilli & Shephard, 1994). There are two main stages in bias detection: statistical detection of differential item performance (DIF) and item review. During item review, items with DIF are carefully examined in order to understand what underlies the differential performance. Finally, an item is considered to be biased if the source of extra difficulty is not relevant to the construct of interest.

DIF detection is becoming an increasingly important part of test validation. In Canada, translation DIF is a particularly important issue for educators, psychologists, and researchers in all

fields because we have two official languages and test takers must be evaluated fairly in their own language (Simon, 1994). Also, policy makers are becoming increasingly interested in large-scale provincial, national, and international assessments in order to compare educational achievement across jurisdictions; use of these tests usually involves translation. Ensuring measurement equivalence in translated tests is quite difficult (Tanzer, 1995). Therefore, studies of DIF are particularly relevant in the Canadian context. When items with substantial DIF are found, test developers should think deeply about the construct that the test is intended to measure and review how well the item measures that construct (Roznowski & Reith, 1999). Systematic application of methods for detecting DIF, corrective action, and subsequent demonstrations that inferences made from tests are unbiased, are vital steps in test validation and will ensure equity in testing (Camilli & Shepard, 1994).

DIF methods for binary data are fairly well developed. However, educational reform has resulted in a movement away from traditional testing with dichotomously scored items toward performance-based assessment (French & Miller, 1996; Zwick, Donoghue, & Grima, 1993). Most performance-based assessments include polytomous items, which provide more information about a student's knowledge than traditional multiple choice items (Chang, Mazzeo, & Roussos, 1996). However, some polytomous items may actually be more likely to measure irrelevant factors and to introduce unfairness in testing (Chang et al., 1996; French & Miller, 1996; Zwick et al., 1993).

Several DIF techniques have been adapted for use with polytomous data. Many are based on extensions of dichotomous methods. However, procedures for assessing polytomous DIF are fairly new and have not yet been well tested. One procedure, Ordinal Logistic Regression (OLR), (Zumbo, 1999) has not been evaluated at all. It is imperative that these techniques are carefully

tested before they are used in applied settings.

The major purpose for this study was to compare four observed score methods for identifying DIF in polytomous items using simulated data which reflect realistic testing conditions. These four methods comprised the Mantel, the Generalized Mantel-Haenseal (GMH), Logistic Discriminant Function Analysis (LDFA) and Ordinal Logistic Regression (OLR). The study will add to the body of knowledge on DIF assessment techniques by assessing a new technique; it will also provide information about the performance of the other techniques under new conditions. Because all four techniques are relatively easy and inexpensive to use, they are practical for applied settings. Thus, results of this study should be a valuable resource for those who develop and use educational tests as well as for measurement theorists.

The second chapter includes a review of relevant literature. DIF is statistically defined, and the leading techniques for identifying DIF in polytomous data are described in detail. Next, empirical studies of DIF detection procedures are presented and summarized. Finally, gaps in knowledge are identified and study objectives outlined.

In the third chapter, the study methods are presented. Details of the measures under study are presented first, and study design presented next. In the third section, steps in data generation and testing of the DIF procedures are outlined. In the fourth section, methods and results from validation of the data generation and DIF procedure programs are presented, while in the fifth section details on the final analyses are presented.

Study results are presented in the fourth chapter. Results are presented separately for each procedure, beginning with Mantel, then GMH, LDFA, and OLR. Each section includes tables and an explanation of the results presented.

In the fifth chapter, strengths and limitations of the study are presented first. Next, results are summarized, discussed, interpreted, and compared to findings of other researchers. Explanations for disparate findings are given.

Finally, in the sixth chapter, conclusions about each of the four procedures are presented. Recommendations for DIF detection in applied settings are given; suggestions for further empirical research are made.

CHAPTER II
LITERATURE REVIEW

In the first section of this chapter, DIF is defined for both dichotomous and polytomous items, and different types of DIF described. In the second section, a description of the leading types of DIF detection procedures for polytomous items is provided. In the third section, studies on polytomous DIF detection procedures are presented; their results are summarized by procedure. In the fourth section, conditions that affect DIF detection are identified and practical issues such as what to do when DIF is found in items are discussed. Finally, gaps in knowledge are identified, study objectives presented, and the contribution of this study to educational measurement highlighted.

Definitions of DIF

DIF refers to a significant difference in the item performance of two groups who are matched on the construct (or ability) measured by the test (Dorans & Holland, 1993). Analyses to estimate DIF compare a reference and a focal group. For example, in the study of a test that was translated from English to French, the English examinees would be the reference group and the French examinees would be the focal group. It is essential to match groups on ability in order to distinguish DIF from real differences between groups (Dorans & Holland, 1993). DIF can be found in both dichotomous items and polytomous items; the definition and presentation of DIF is more complicated in polytomous items.

DIF in dichotomous items

In dichotomous items, DIF is a difference in the probability of correct response between two groups who are equal in ability. DIF assessment procedures for dichotomous items can be classified according to whether they use observed score or latent variable method as the matching variable (Potenza & Dorans, 1995). Observed score methods test whether the proportion of examinees getting the item correct is the same in both groups, after they are matched on total test score. Chang and his colleagues (1996) provided statistical definitions of DIF for both observed score methods and latent variable methods. Methods based on total test score as the matching variable provide a test of whether

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

holds for all values of X, when Y is the score of the studied item, X is the total test score, and $E_R[Y|X]$ and $E_F[Y|X]$ are the regressions of the item score Y on X (Chang et al, 1996).

Latent variable methods divide observed test score into two parts: a reliable portion and an unreliable portion (Potenza & Dorans, 1995). The reliable portion is called the latent trait or θ. Latent trait methods test whether there is a significant difference in item score for two groups who have equal levels of θ or, in statistical terms, whether

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

for all values of θ when $E_R[Y|\theta]$ and $E_F[Y|\theta]$ are the regressions of the item score Y on θ (Chang et al., 1996).

DIF in dichotomous items can be uniform or nonuniform. Uniform DIF occurs when the probability of correctly answering an item is greater for one group than for the other across the entire ability range; this means that the item is consistently more difficult for one group than for

the other (Zumbo, 1999). Nonuniform DIF occurs when there is an interaction between ability level and group membership; the probability of answering the item correctly is higher for the reference group at some points on the ability scale, and higher for the focal group at other points on the ability scale (Zumbo, 1999). Nonuniform DIF means that the item has higher discrimination (relation to ability) for one group than for the other.

DIF Assessment Procedures for Dichotomous Items

Many DIF assessment procedures are available for dichotomous items. The most popular include: SIBTEST (Shealy & Stout, 1993), logistic regression (Swaminathan & Rogers, 1990), procedures based on item response theory (Thissen, Steinberg, & Wainer, 1988) the Mantel-Haenszel procedure (Holland & Thayer, 1988), and a descriptive index, called standardization (Dorans & Kulick, 1986). All are 'internal methods' which are designed to test whether items measure the construct assessed by the test equally well for both groups (Dorans & Holland, 1993).

DIF in Polytomous Items

Polytomous items have more than two score categories; these are ordered from worst to best (Zwick et al, 1993). Many of the performance items in the Third International Mathematics and Science Study are polytomous (Harmon, Smith, Martin et al, 1997); the second part of Item 1 from the containers task is presented in Appendix One as an example. Item 1, Part 2 has a maximum score of 3 and a total of four score levels (0,1,2, and 3) (Harmon, Smith, Martin et al, 1997). Thus, there are three score thresholds; one for the probability of getting 0 rather than 1, one for the probability of getting 1 rather than 2, and one for the probability of getting 2 rather than 3.

In polytomous items, DIF is a significant difference in item score thresholds between the two groups matched on ability (either total test score or latent ability). DIF assessment procedures for polytomous items can also be classified into observed score and latent trait approaches (Potenza & Dorans, 1995).

When total score is used as the matching variable, DIF is defined as a significant difference between the reference and focal group in the regressions of the polytomous item score thresholds on the total test score (Potenza & Dorans, 1995). The studied item must be included in the matching variable to lower the probability of Type I Error (Zwick et al, 1993).

When latent ability is used as the matching variable, DIF is defined as a significant difference between the reference and focal groups in the regression of the polytomous item scores on the latent matching variable (Potenza & Dorans, 1995; Zumbo, 1999). In statistical terms, the regression of item score on ability can be defined as a weighted sum of item category response functions.

$$E_g[Y|\theta] = \sum_{k=1}^{m} kP_{k.g}(\theta) \tag{3}$$

where Y is scored in terms of $m + 1$ ordered categories ($Y = k$, $0 \leq k \leq m$) (Chang et al, 1996) and $P_{kg}(\theta)$ denotes the item category response function, or the probability that a randomly selected examinee with ability θ from group g (g = R for the reference group and F for the focal group) will receive item score k.

Null DIF is defined as:

$$P_{kR}(\theta) = P_{kF}(\theta),\ k = 1,....m \tag{4}$$

An item does not have DIF if the regression of polytomous item scores on the latent variable is identical for the reference and focal groups (Chang et al, 1996).

Uniform and nonuniform DIF occur in polytomous items as well as in dichotomous items; uniform DIF is the most common. The results of several applied studies are presented in Table One. This table shows that both uniform and nonuniform DIF do occur in real test data, and that the level of DIF found in some tests is rather high.

When uniform DIF occurs in polytomous items, the item is consistently more difficult for one group than another across all score levels (French & Miller, 1996). Figure 1 shows what uniform DIF might look like in an item which is polytomously scored from 0 to 3[1]. It can be seen

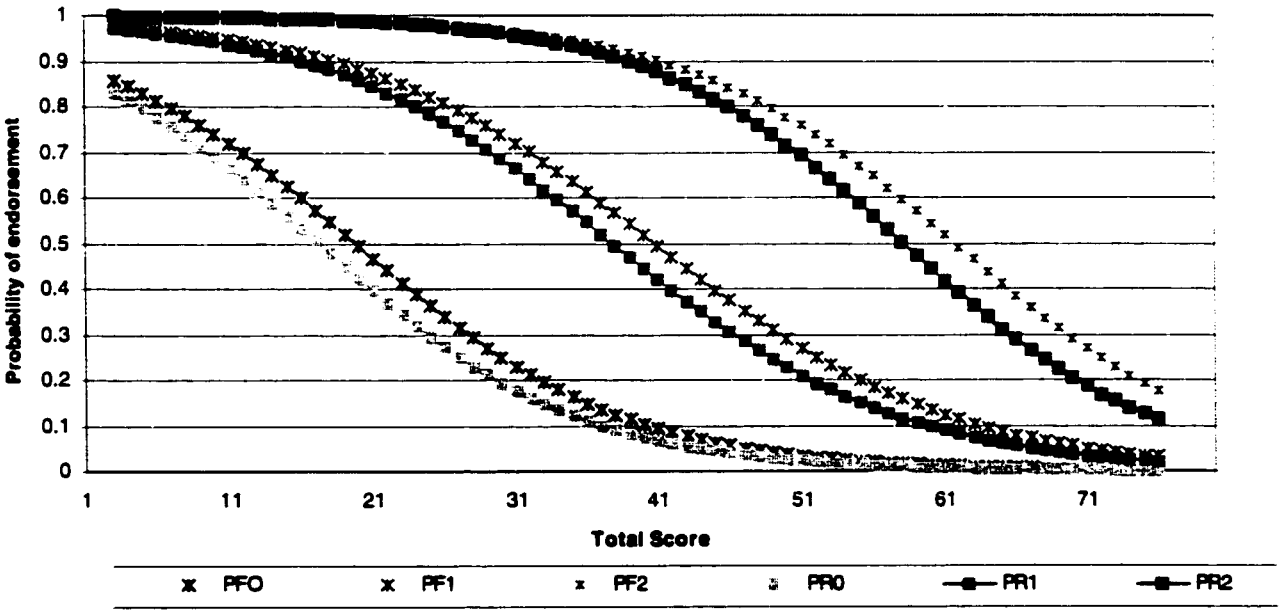


Figure 1. An example of uniform DIF in a polytomous item with four score levels and three score thresholds.

[1] Please note that this graph is based on simulated data rather than on a real item.

that, for all three score thresholds (probability of getting a 0 versus that of getting a 1, of 1 versus that of 2, and that of 2 versus 3), the focal group has a consistently higher probability of receiving the lower score than the reference group (with no DIF, all lines for the focal and reference group would overlap) .

Non-uniform DIF in polytomous items is characterized by a group by ability interaction; the item does not discriminate equally well for the reference and focal groups (French & Miller, 1996; Zumbo, 1999). Figure 2 depicts an item with nonuniform DIF. Here, for each threshold, when ability is low, the reference group is more likely to have lower scores; when ability is high, the focal group is more likely to have lower scores.

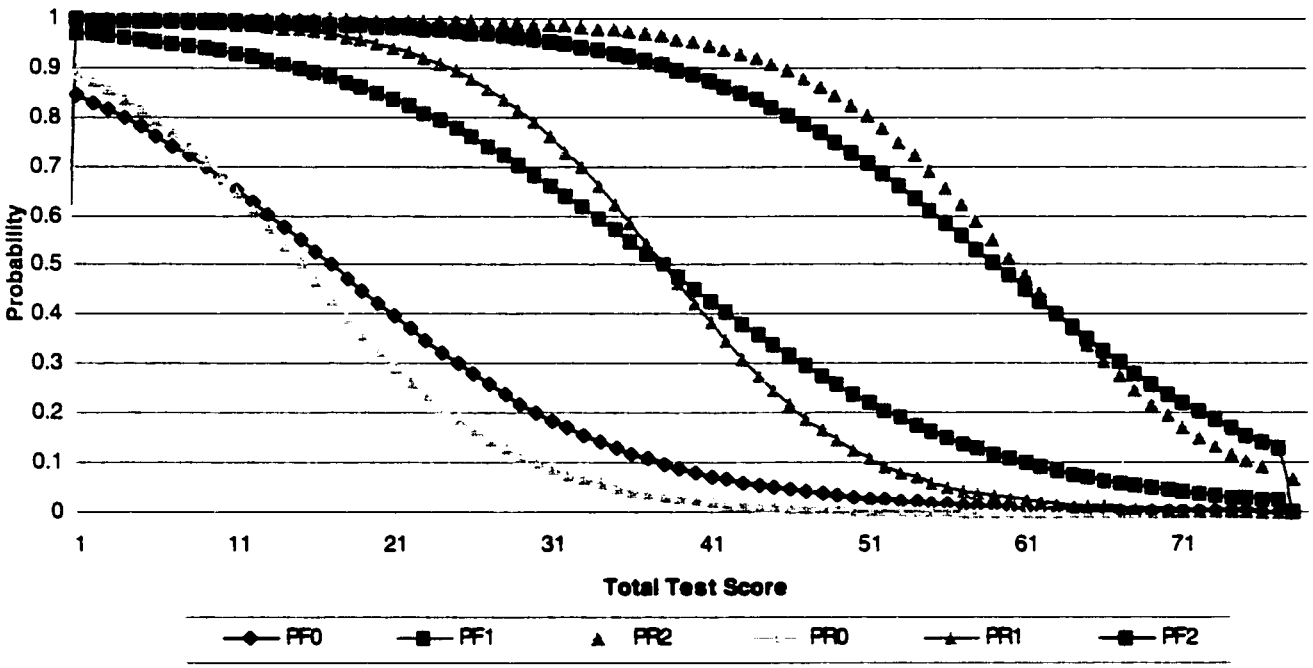


Figure 2. An example of nonuniform DIF in a polytomous item with four score levels and three score thresholds.

Another type of DIF, called balanced DIF, has been simulated in empirical studies of DIF detection procedures. In balanced DIF, the reference group would be favored at one score threshold, while the focal group would be favored at another score threshold. Although this type of DIF has been simulated in several empirical studies (e.g. Zwick et al., 1997; Tian, 1999), it is extremely unlikely to occur in real data (Ankenmann, Dunbar, & Dewitt, 1999; Spray, personal communication, June 30, 1999).

Table 1. DIF Found in Actual Test Data

Study	Type of item	Method used	Type of DIF (Magnitude if available)
Zwick et al, 1997, American Physics Board Exam	6 Free response items	STND, Mantel, STND-H, STND-M	Mean predicted difference after total score controlled averaged 0.51 out of 15 points.
Bjorner et al, 1998, Danish translation of the SF-36	36 items, all polytomous. Multidimensional instrument	Gamma an χ^2	Six items had slight to moderate DIF (10-15 % differences in item scores); 5 items had significant DIF (15% or more difference in item scores). Most had uniform DIF; 2 had both.
Marshall et al, 1997 Min-Mental State Exam in Spanish	Most dichotomous; 1 polyomous	Logistic regression	2 items uniform only, 2 nonuniform, 4 had both, 3 unrelated to test score (11 DIF items contributed to a mean difference of 3 points out of total test score of 30).
Smith and Reise, 1998 Stress Reaction Scale Multidimensional Personality Scale			Adjusted differences in *b* ranged from 0.19 to 0.65.
Ellis, 1989 American and German intelligence tests	Dichotomous	IRT. Lord's χ^2	8 of 106 in American test had DIF; 2 of 145 in German test had DIF. 3 had uniform DIF, 5 nonuniform, and 2 had both. .
Budgell et al, 1995. French translations of government aptitude tests			Odds ratios ranged from 1.37 to 2.39. Didn't discriminate between uniform and nonuniform.
Price et al, 1998, Japanese translation of diving exam (mastery)	30 item test, dichotomous		6 out of 30 uniform DIF; 4 had nonuniform DIF

DIF Assessment Procedures for Polytomous Items

Potenza and Dorans (1995) have outlined a two-dimensional scheme for classifying DIF assessment procedures. Procedures are classified on the first dimension according to whether total test score or latent ability is used as the matching variable and on the second dimension as parametric or non-parametric. DIF assessment procedures may be further classified as descriptive or inferential; some of the newer procedures include both a descriptive and an inferential statistic.

The validity of DIF assessment depends on the quality of the matching variable; methods which use latent trait as the matching variable are theoretically preferable because they correct for unreliability in the test score. However, Zwick and her colleagues (1993) have shown that there is a theoretical justification for using the total score if the data are unidimensional, and if the data follow the Partial Credit or General Partial Credit Model, and if the studied item is included in the total test score. If the number of items in the total score is too small, or if the studied item is not included in the total score, then DIF assessment is problematic.

In the following sections, DIF assessment procedures are ordered using the classification used by Potenza and Dorans (1995); each is described in detail.

Total Test Score Methods: Non-Parametric

These methods include the Mantel (Mantel, 1963; Zwick et al, 1993), the Generalized Mantel-Haenszel (Mantel & Haenszel, 1959; Zwick et al, 1993), the HW1 and HW3 (Welch & Hoover, 1993), and the Standardized expected item score Mean Difference, or $STND_{ES\text{-}DIF}$ (Dorans & Schmitt, 1991; Dorans, Schmitt, & Bleinstein, 1992). The $STND_{ES\text{-}DIF}$ is a descriptive procedure while the others are inferential procedures.

These are all extensions of the Mantel-Haenszel procedure (Mantel & Haenszel, 1959), which is one of the most popular and commonly used procedures for detecting DIF in dichotomous items (Potenza & Dorans, 1995). The Mantel-Haenszel is a contingency table procedure in which the data for the studied item for the reference and focal groups are arranged into a series (based on each level of total test score) of two by two-contingency tables of counts of right and wrong answers. The null hypothesis is that the odds of getting the item correct are the same for the focal and reference group across all levels of the matching variable (Potenza & Dorans, 1995). Holland & Thayer (1986) derived a descriptive measure of effect size (amount of DIF) using a log-odds transformation of the odds-ratio (α MH) into a difference on the delta scale called $MH_{D\text{-}DIF}$ (Potenza & Dorans, 1995). $MH_{D\text{-}DIF}$ is used at Educational Testing Services to classify items by the level of DIF shown: Category A, or No DIF, category B, or slight DIF, and Category C, or important DIF. Items in Category C have significant $MH_{D\text{-}DIF}$ and an absolute value of $MH_{D\text{-}DIF}$ exceeding 1.5 (roughly equal to a 15% difference in item score between groups) (Zieky, 1993).

The Mantel

Mantel (1963) proposed an extension of the MH procedure for ordered response categories that involves comparing the means between matched groups. Zwick and her colleagues (1993) provided equations for using the Mantel as a DIF detection procedure for polytomous items. As with the dichotomous MH procedure, the reference and focal groups are first matched on ability, or total test score. Next, the data are arranged into a 2 by T by K contingency tables where T is the number of response categories and K is the number of levels of the matching variable (Zwick et al, 1993).

Table 2. Sample Contingency Table (from Zwick et al., 1993)

	Item Score				
	y_1	y_2	y_3	y_T	*Total*
Focal	n_{F1k}	n_{F2k}	n_{F3k}	n_{FTk}	n_{F+k}
Reference	n_{R1k}	n_{R2k}	n_{R3k}	n_{Rtk}	n_{R+k}
Total	n_{+1k}	n_{+2k}	n_{+3k}	n_{+Tk}	n_{++k}

Note: This table is repeated for each level of total score

where:

n_{++k} is the number of examinees at the kth level of total score (matching variable) and

n_{F+k} and n_{R+k} are the total number of examinees in the focal and reference group, respectively, at the kth level of total score.

$y_1, y_2, y_3 \ldots y_k$ are the T scores that can be obtained on the item, and

n_{Fik} and n_{Rik} refer to the number of focal and reference group members at the kth level of the matching variable who received an item score of y_T.

Calculation of the Mantel is based on a comparison of the item means for matched groups. The null hypothesis is that at a fixed level of total score, there is no conditional association between the item score and group membership (Potenza & Dorans, 1995). The following formula is used:

$$Mantel\ \chi^2 = \frac{\left(\sum_k F_k - \sum_k E(F_k)\right)^2}{\sum_k Var(F_k)} \tag{5}$$

where:

$$F_k = \sum_t y_t n_{Ftk} \sum Y_k N_{Fmk} \tag{6}$$

(7)

and $$E(F_k) = \frac{n_{F+k}}{n_{++k}} \sum_t y_t n_{+tk}$$

$$Var(Fk) = \frac{N_{R+k} N_{F+k}}{N^2{}_k (N_{++k} - 1)} \left\{ \left(N_{++k} \sum_t Y^2{}_t N_{++k} \right) - \left(\sum_t Y_t N_{+tk} \right)^2 \right\} \tag{8}$$

Rejecting the null hypothesis means that members of the reference and focal group who have equal ability (same test score) differ in their mean item score (Potenza & Dorans, 1995).

The GMH

The GMH is a generalized MH statistic for nominal response data that is based on group differences in the entire response distribution (Mantel & Haenszel, 1959). The null hypothesis is that there is no association between item responses and group membership for a given level of total test score (Zwick et al, 1993).

The GMH χ^2 test statistic is given by:

$$\text{GMH } \chi^2 = \left[\sum A_k - \sum E(A_k)\right]'\left[\sum V(A_k)\right]^{-1}\left[\sum A_k - \sum E(Ak)\right] \quad (9)$$

Where:

$$A_k = (n_{R1k}, n_{R2k}, \ldots n_{R(T-1)k}) \quad (10)$$

$$E(A_k) = n_{R+k}\mathbf{n}_k / n_{++k} \quad (11)$$

$$\mathbf{n}'_k = (n_{+1k}, n_{+2k}, \ldots \mathbf{n}_{+(T-1)k}) \quad (12)$$

$$V(A_k) = n_{R+k}n_{F+k}\left(\frac{n_{++k}\,diag(n_k) - n_k n_k'}{n^2_{++k}(n_{++k}-1)}\right) \quad (13)$$

dia (n_k) is a (T-1) by (T-1) matrix with elements n_k.

A'_k and $E(A'_k)$ are vectors of length T-*1* and

$V(A_k)$ is a (T-1) by (T-1) covariance matrix.

The GMH statistic is chi-square, with T-1 degrees of freedom.

The Standardized Expected Score Mean Difference

The $STND_{ES\text{-}DIF}$ (Dorans & Schmitt, 1991, Dorans et al, 1992) is a descriptive statistic which shows the magnitude of DIF; it can also be converted into a *z*-statistic for hypothesis testing (Zwick & Thayer, 1996). $STND_{ES\text{-}DIF}$ is an extension of the dichotomous standardization statistic $STND_{P\text{-}DIF}$ developed by Dorans & Kulick, 1986 (Potenza & Dorans, 1995). The statistic compares the mean of the reference and focal groups at each level of total score; this difference is weighted by focal group frequencies to give the reference group and focal group the same distribution across levels of the matching variable (Zwick et al., 1993). This gives more weight to score levels with a higher frequency and makes the index sensitive to the total scores obtained most often by focal group examinees (Camilli & Shepard, 1994).

To calculate $STND_{ES\text{-}DIF}$, let *X* be the matching variable with *M* levels, $m = 1,....,M$, and let *Y* be the ordered item score with *K* categories, $k = 1,....,K$. First, expected item scores for the reference ($E_{Rm}(Y|X)$) and focal ($E_{Rm}(Y|X)$) groups are calculated (Potenza & Dorans) using:

$$E_{Fm}(Y|X) = \Sigma N_{Fmk} Y_k / N_{Fm} \quad \text{and} \tag{14}$$

$$E_{Rm}(Y|X) = \Sigma N_{Rmk} Y_k / N_{Rm} \tag{15}$$

where:

N_{Fmk} is the number of examinees in the focal group at score level *m* with item score Y_k,

N_{Fm} is the total number of examinees in the focal group at score level *m*

N_{Rmk} is the number of examinees in the reference group at score level *m* with item score Y_k

and N_{Rm} is the total number of examinees in the reference group at score level *m*

The item score, Y_k can take on any ordered values, including 1,2,3...*K*

Then, the differences in expected item score at each level of the matching variable are computed:

$$D_m = E_{Fm}(Y|X) - E_{Rm}(Y|X) \quad (16)$$

Next, these differences are weighted by focal group relative frequencies:

$$STND_{ES\text{-}DIF} = \sum_m F_m D_m / N_F \quad (17)$$

where N_F is the total number of focal group examinees. A negative $STND_{ES\text{-}DIF}$ means that, after matching on total score, the focal group has a lower mean score on the item (Zwick et al., 1993). $STND_{ES\text{-}DIF}$ may be used as a descriptive supplement to the Mantel.

More recently, Zwick and Thayer (1996) derived two inferential tests based on the $STND_{ES\text{-}DIF}$. The SMD-H and SMD-M are calculated by dividing the $STND_{ES\text{-}DIF}$ by standard errors derived under the hypergeometric model and multinomial model respectively (Zwick & Thayer, 1996).

HW1 and HW3

HW1 and HW3 (Welch & Hoover, 1993) are two inferential statistics which fall into the standardization framework. In both procedures, differences in expected item score are averaged across levels using weights (Potenza & Dorans, 1995). For HW1, the difference in expected item score at each level is calculated and converted into a *t* statistic by dividing by a pooled standard error of the mean difference (Potenza & Dorans, 1995). These *t* statistics are summed across levels and divided by the square root of the sum of variances of the *t* statistics (Welch & Hoover, 1983). HW1 is normally distributed with a mean of zero and a variance of one. In HW3, the test

statistic is weighted by the reciprocal of its sampling variance; like HW1, it is normally distributed with a mean of 0 and a standard deviation of one (Potenza & Dorans, 1995).

Total Test Score Methods: Parametric

Ordinal Logistic Regression (Zumbo, 1999) and logistic discriminant function analysis (LDFA; Spray & Miller, 1994) are both parametric, model-based methods which use total test score as the matching variable. In both, uniform and nonuniform DIF can be modeled with the same equation and separate tests for the two types of DIF are provided.

Ordinal Logistic Regression

In the dichotomous logistic regression procedure, the probability of observing each dichotomous response, U, is modelled. Three explanatory variables are used: observed test score, X, a group indicator variable, G, and the interaction between G and X. The dichotomous logistic regression model for DIF can be given by the following equation for each group:

$$P(u_{ij} = 1 \mid \theta_{ij}) = \frac{e^{(\beta_{0j} + \beta_{1j}\theta_{1j})}}{[1 + e^{(\beta_{0j} + \beta_{1j}\theta_{1j})}]} \qquad (18)$$

where:

u_{ij} = the response of person i in group j to the item

θ_{ij} = the observed ability of individual i in group j

β_{0j} is the intercept parameter, and

β_{1j} is the slope parameter for group j.

No DIF is present if the logistic regression curves for the two groups are the same; that is, if

$\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$. If $\beta_{01} \neq \beta_{02}$ but $\beta_{11} = \beta_{12}$ then the curves are parallel and do not cross; thus the item shows uniform DIF (Swaminathan & Rogers, 1990). If $\beta_{01} = \beta_{02}$ but $\beta_{11} \neq \beta_{12}$ the curves are non- parallel and the item shows non-uniform DIF (Swaminathan & Rogers, 1990).

Logistic regression has been used successfully for detecting DIF in dichotomous items. In general, logistic regression displays good Type I Error control as well as high power for detecting both uniform and nonuniform DIF (Swaminathan & Rogers, 1990).

French and Miller (1996) suggested that logistic regression might be useful for DIF detection in polytomous data. There are many extensions of logistic regression to accommodate polytomous data (Agresti, 1990); most involve recoding polytomous data into a number of dichotomous sets (J - 1 where J is the number of score levels). French and Miller (1996) described three such approaches to coding polytomous data for logistic regression: continuation logits, cumulative logits, and adjacent categories logits. All three approaches are based on the logit, or the ratio of the probability of getting one score on the item to that of getting a score that is one higher or one lower. An item with four score categories would have three logits, and three logistic curves. In all three approaches, separate regressions are run for each logit in every model (nonuniform, uniform, and null) that is tested. For example, testing the uniform, nonuniform, and null DIF models for a four-point answer scale would require nine regressions; three for each logit * three models. The three approaches to coding are briefly described below, again using the example of a four-point response scale.

The continuation ratio logits approach uses an increasingly narrow amount of data in each successive regression to test for the presence of DIF. In the first regression, the probability of receiving a score of zero is compared to the probability of receiving a score greater than zero. In

the second regression, the probability of receiving a one is compared to that of receiving a score greater than one, and in the third regression, the probability of receiving a two is compared to the probability of receiving a three. An overall χ^2 test is available for this method of coding.

In the adjacent categories logit model, each response probability is compared to the response probability adjacent to it. In the first regression, those that received a score of zero are compared to those who received a score of one, in the second, those who receive a score of one are compared to those who receive a score of two, and in the third, those who receive a score of two are compared to those who receive a score of three. Because only two categories are compared in each regression, much information is lost in this coding scheme (French & Miller, 1996).

With cumulative logits coding, in the first regression the probability of receiving a 0 is compared to the probability of receiving any other scores. In the second regression, the probability of receiving one is compared to that of receiving any other scores, and in the third regression, the probability of receiving a 2 is compared to the probability of receiving all other scores. The cumulative logits model is theoretically and empirically preferable to the other two coding methods because no information is lost in coding (French & Miller, 1996).

French and Miller (1996) tested these three coding schemes as methods for detecting DIF in polytomous items. They found that although the cumulative and continuiation-ratio logits coding worked fairly well, running separate regressions for each logit was unwieldy. Furthermore, they found that it was difficult to come up with overall conclusions about the presence of DIF based on J - 1 separate regressions.

The OLR Procedure. Zumbo (1999) built on the work of Swaminathan and Rogers (1990) and French and Miller (1996) to develop Ordinal Logistic Regression as a procedure for DIF assessment. Ordinal Logistic Regression is a relatively new addition to statistical computer packages such as SAS and SPSS. In Ordinal Logistic Regression, the cumulative logits coding scheme is used, but all logits are tested simultaneously. Thus, OLR overcomes the difficulties of multiple testing and interpretation that were noted by French and Miller (1996).

The OLR DIF detection procedure outlined by Zumbo includes a measure of effect size in addition to the statistical test. In the full model, the linear regression of predictor variables on an unobservable continuously distributed random variable, y* is tested, where:

$$y^* = b_0 + b_1 \text{ TOT} + b_2 \text{ GROUP} + b_3 \text{ TOT} * \text{GROUP} + \xi_i. \quad (19)$$

This means that y* is a function of the intercept, total test score (TOT), group membership (GROUP), an interaction between total score and group, and error. An appropriate equation for DIF analyses is:

$$\text{logit}\,[(P(Y \leq j)] = \alpha\, j + b_1 tot + b_2 group + b_3\,(tot * group). \quad (20)$$

This is the full model, which models both uniform and nonuniform DIF. The model for uniform DIF is expressed as

$$\text{logit}\,[(P(Y \leq j)] = \alpha\, j + b_1 tot + b_2 group. \quad (21)$$

The model for null DIF can be expressed as:

$$\text{logit}\,[(P(Y \leq j)] = \alpha\, j + b_1 tot. \quad (22)$$

Unlike the procedures described by French and Miller, this procedure provides a simultaneous test of all cumulative logits in one regression. However, the logits can only be tested simultaneously if the assumption of equal slopes for all logistic curves is upheld. If this

assumption does not hold, each logit must be tested with a separate regression in each model. Thus, with a four-point response scale, testing each of the three models would require nine regressions (three for each).

If the assumption of equal slopes is valid, testing for the significance and magnitude of DIF involves three steps (Zumbo, 1999). In the first step, three regressions are run, one for the null model, one for the uniform DIF model, and one for the total DIF (uniform and nonuniform) model. At each step, a χ^2 and an R^2 value are provided as output. In Step Two, the difference in χ^2 and R^2 values between the total DIF model and the null model are calculated in order to test for the presence of DIF of any type (either nonuniform or uniform). DIF is present if the difference in χ^2 at two degrees of freedom is significant, and if the R^2 exceeds a specified amount (Zumbo suggests 0.13). Step Three is only performed if there is DIF; it involves the comparison of χ^2 and R^2 values from the model for uniform DIF to those from the fully saturated model in order to determine whether the DIF is uniform, nonuniform, or both.

The Ordinal Logistic Regression procedure presented by Zumbo (1999) is both conceptually sound and practical for applied situations: 1) Decisions about DIF are based on a combination of effect size and statistical significance. This should ensure that important effects are not hidden by small sample sizes and that trivial effects are not deemed statistically significant. 2) It is model based, and can provide a test of overall DIF as well as separate tests of uniform and nonuniform DIF. 3) It is computationally efficient, especially if the assumption of equal slopes holds 4) Finally, the sequential test strategy outlined by Zumbo in which a test of overall DIF in the first step is conceptually sound. The first priority in DIF assessment should be

to determine whether DIF of any kind is present. After this, classification of the DIF is relevant.

The basic OLR procedure is available in SPSS; an SPSS macro (Kuehnel, 1999) is used to derive the R^2 measure of effect size.

LDFA

Miller and Spray (1993) demonstrated that logistic discriminant function analysis (LDFA) may be used to detect DIF in polytomous items. In a DIF context, LDFA can be used to classify people into groups that they most closely resemble on the basis of item responses and test score. According to Miller and Spray (1996), LDFA has several advantages: 1) it is model-based and so provides tests for both uniform and nonuniform DIF, 2) it requires no assumption of normality, and 3) it is much easier to implement than polytomous logistic regression because only one regression is needed to test each model (Miller & Spray, 1993).

LDFA is closely related to polytomous logistic regression and log-linear analysis; one difference is that the regression models in LDFA are used to predict group membership rather than item score. The LDFA procedure involves a sequential test strategy for nonuniform and uniform DIF utilizing three models: nonuniform, uniform, and null DIF. Nonuniform DIF is modelled in Equation 24. In it, the probability of group membership is predicted from total score, item response, and the interaction between item response and total test score (Spray & Miller, 1994).

$$\text{Prob}\,(G \mid X,U) = \frac{e^{(1-G)(-\alpha_0 - \alpha_1 X - \alpha_2 U - \alpha_3 X * U)}}{1 - e^{(-\alpha_0 - \alpha_1 X - \alpha_2 U - \alpha_3 X * U)}} \qquad (23)$$

where:

α_i = 0,1,2,3 are the discriminant function coefficients to be estimated

G = group indicator variable; 1 = reference and 0 = focal

X = observed test score and:

U = item response variable that can take on any one of J values and

Pr = probability of group membership for the reference or focal group

This model is very similar to the one given in Equation 19. However, there are some differences: 1) group membership (G) is modelled instead of item response (U), 2) item response is not restricted to two categories but can take on any value, 3) item response (U) is a predictor rather than the dependent variable, and 4) the regression coefficients are given by α_i = 0,1,2,3 to distinguish them from the β coefficients in Equation 19.

Uniform DIF is modelled in Equation 25. Here, the probability of group membership is predicted from total score and item response.

$$\text{Prob}\,(G|\,X,U)=\frac{e^{(1-G)(-\alpha_0-\alpha_1 X-\alpha_2 U)}}{1-e^{(-\alpha_0-\alpha_1 X-\alpha_2 U)}} \tag{24}$$

Equation 26 represents null DIF. Here, the probability of group membership is modelled from total score only.

$$\text{Prob}\,(G|\,X,U)=\text{Prob}\,(G|\,X)=\frac{e^{(1-G)(-\alpha_0-\alpha_1 X)}}{1-e^{(-\alpha_0-\alpha_1 X)}} \tag{25}$$

A likelihood ratio goodness-of-fit statistic,G^2, can be computed for each model. The difference in G^2 values between that from the model given in Equation 24 and that from the uniform DIF model in Equation 25 provides a test of nonuniform DIF; this difference is

distributed as χ^2 with one degree of freedom. The difference in G^2 values between the uniform DIF model given in Equation 25 and that of the null model in Equation 26 provides a test of uniform DIF; it is also distributed as χ^2 with one degree of freedom. If the test for either 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 from that which would be predicted from test score alone.

Latent Variable Methods: Parametric

Parametric latent trait procedures for assessing DIF are all based on item response theory (Potenza & Dorans, 1995). With dichotomous data, the item characteristic curve depicts the regression of probability of a correct response on the latent variable of interest (ability). The *y* intercept is the probability that someone with very low aptitude will get the item correct (Zumbo, 1999). The position along the *x* axis indicates the amount of the latent trait needed to get the item correct (item difficulty, *b* parameter). The slope (*a* parameter) indicates how well the item discriminates among examinees of different ability (Zumbo, 1999). An item is said to be functioning differentially when the probability of correct response to the item is different for examinees of the same ability level, but from different groups (Kim & Cohen, 1998). IRT based DIF assessment methods for dichotomous data include the general IRT likelihood ratio, IRT^{D2}, loglinear IRT-LR, and Lord's χ^2 (Potenza & Dorans, 1995).

In polytomous IRT models, if all sets of item parameters in two groups are equal, item response functions will be equal. An item is said to display DIF if the item response functions in the reference and focal groups are unequal (Kim & Cohen, 1998). Several different approaches for testing the equality of item parameters in different groups have been developed, including

direct comparison of item parameters from different groups, comparison of areas between expected item scores, Muraki's General Partial Credit Model (Muraki, 1993), and comparison of likelihood functions to evaluate the differences between item responses in the two groups (LRT) (Kim & Cohen, 1998). Theoretically, IRT models are preferable for DIF detection because they actually model what is being tested. However, they are impractical for applied settings for several reasons: 1) Strict assumptions of model-data fit and normality of the underlying ability distributions must be met (Potenza & Dorans, 1995) 2) All IRT DIF procedures require a minimum sample size of 500 in each group for correct item parameter estimation (Kim & Cohen, 1998) 3) They are computationally very expensive (Potenza & Dorans, 1995) 4) IRT procedures require purchase of special computer programs (Potenza & Dorans, 1995). For these reasons, IRT DIF procedures are not reviewed further in this paper.

Latent Nonparametric Procedures

SIBTEST

SIBTEST is one of the better techniques for the assessment of DIF in dichotomous items (Potenza & Dorans, 1995). SIBTEST is based on Stout's theory of item and test performance in which the target trait (purpose of test) is distinguished from secondary nuisance traits; DIF results from unequal distribution of the nuisance traits between focal and reference groups (Chang et al, 1996; Potenza & Dorans, 1995). The amount of DIF for a given level of latent ability (true score) is given by Chang et al.,1996 as:

$$B_0(\theta) \equiv E_r[Y|\theta] - E_F[Y|\theta]. \tag{26}$$

This measure parallels the $STND_{P\text{-}DIF}$; the differences in the item-score regressions are averaged across levels, with focal group weighting. However, the ability estimate in SIBTEST is one based

on a correction for unreliability in test score; this improves the matching variable (Potenza & Dorans, 1995). Shealy and Stout (1993) developed a global index of DIF for the dichotomous case:

$$\beta = \int B_0(\theta) f_F(\theta) d\theta \tag{27}$$

This global index β can be interpreted as the expected amount of DIF experienced by a randomly selected focal group examinee. The null hypothesis tests $\beta = 0$ versus $H_1: \beta \neq 0$. The SIBTEST procedure includes a descriptive index of the amount of DIF in an item as well as a test of significance. SIBTEST only detects uniform DIF, but Li and Stout (1996) later adapted SIBTEST to detect crossing DIF.

Chang and his colleagues (1996) have adapted SIBTEST to study polytomous uniform DIF using the classical test theory latent variable definition of DIF. In Poly-SIBTEST (Chang et al, 1996) the local measure of DIF at the matching true test score t can be defined as:

$$B(t) = E_R[Y|t] - E_F[Y|t] \tag{28}$$

and the corresponding descriptive DIF index can be defined as

$$\beta = \int B(t) f_F(t) dt \,. \tag{29}$$

DIF can be estimated by: $d_k = \overline{Y}_{Rk} - \overline{Y}_{Fk}, k = 0, \ldots n_h$ where (30)

$\overline{Y}_{Rk} - \overline{Y}_{Fk}$ is the group difference in performance on the studied item among examinees with the same test score.

The following statistic can be used to estimate the DIF β in the case when reference and focal groups have equal ability distributions:

$$\beta = \sum_{k=0}^{n_k} p_k d_k \tag{31}$$

and a test statistic can be defined as:

$$B = \frac{\hat{\beta}}{\hat{\sigma}(\hat{\beta})} \tag{32}$$

where

$$\hat{\sigma}(\hat{\beta}) = \left[\sum_{k=0}^{n_H} p_k^2 \left(\frac{\hat{\sigma}^2(Y|k,R)}{N_{Rk}} + \frac{\hat{\sigma}^2(Y|k,F)}{Nk} \right) \right]^{1/2} \tag{33}$$

and

$$\hat{\sigma}^2(Y|m,G) \tag{34}$$

is the sample variance of the studied item scores for examinees in group with total test score X = k on the matching test. However, if examinees are matched on total score, type I Error is likely to be inflated in cases where there are group distributions in ability. Therefore, the matching true test scores for Y_{Rk} and Y_{Fk} are estimated by using a linear regression of true score on observed score where Chronbach's alpha is used as the slope of the regression line. Thus, the actual statistic used in the modified SIBTEST for polytomous items includes a correction to control for Type I Error due to differences in group ability (Chang et al, 1996). It is:

$$\hat{\beta}^* = \sum_{k=0}^{n} p_k d_k^* \tag{34}$$

where adjusted d_k^* s are calculated by the linear regression described above.

Empirical Studies of Polytomous DIF Assessment Procedures

In the following section, the empirical evidence from Monte Carlo studies on the validity of seven DIF assessment techniques for polytomous items is summarized. This review includes studies of the Mantel (Mantel, 1963), $STND_{ES\text{-}DIF}$ (Dorans & Kulick, 1986), the GMH (Mantel & Haenszel, 1959), HW1 and HW3 (Welch & Hoover, 1993), LDFA (Spray & Miller, 1993), Polytomous Logistic Regression (French & Miller, 1996), and poly-SIBTEST (Chang, 1996). Ordinal Logistic Regression (Zumbo, 1999) has not yet been studied.

After a thorough search of the educational and psychological measurement literature, nine relevant Monte Carlo studies were identified: Zwick, Donoghue, and Grima (1993), Zwick, Thayer, and Mazzeo (1997), Welch and Hoover (1993), Spray and Miller (1994), Chang, Mazzeo, and Roussos (1996), French and Miller (1996), Tian (1999), and Ankenmann, Witt, and Dunbar (1999).

Zwick, Donoghue, and Grima (1993) compared the Type I Error and power of the Mantel and GMH for uniform, balanced, low-shift (DIF that affected only the low score categories), and high-shift (DIF that affected only the high score categories); they also studied the performance of the $STND_{ES\text{-}DIF}$ in showing the magnitude of DIF by studying the mean and standard deviation of that statistic. In their simulated 25-item test, twenty items were dichotomous and five were polytomous with four score levels; the 25th (polytomous) item was the studied item. DIF magnitude (difference between *b*-parameters of 0.1 or 0.25), group ability distribution (equal and unequal with a difference of 1.0 standard deviations), and method for matching (include the studied item or not) were varied. Sample size was 500 in each group. They found that while the Mantel had higher power than the GMH for detecting uniform DIF, the GMH had much higher

power for balanced DIF. The $STND_{ES\text{-}DIF}$ was only sensitive to uniform DIF. Group ability differences resulted in slightly increased Type I Error rate for all procedures. However, the mean Type I Error rate for the Mantel, GMH, was lower than 0.05 across all conditions. They also found that the studied item should be included in the matching test to reduce Type I Error.

In a later study, Zwick, Thayer, and Mazzeo (1997) compared the Type I Error and power for uniform DIF of three descriptive and five inferential procedures for DIF detection in polytomous items. The descriptive procedures included the $STND_{ES\text{-}DIF}$ and two procedures based on poly-SIBTEST while the inferential procedures included Mantel, two based on $STND_{ES\text{-}DIF}$ divided by standard errors (one hypergeometric and one multinomial), and two based on poly-SIBTEST. Sample size was 500 in each group, and all matching test items were dichotomous with no DIF; the studied item had four score levels. Three levels of studied item discrimination (0.47, 0.86, and 1.57) were used. The effect of varying the overall item difficulty parameters for the reference group (- 0.5 and 0.5) was also studied as was the effect of differences in overall reference and focal group item difficulties (- 0.25, 0, and 0.25). Zwick, Thayer, and Mazzeo (1997) found a significant interaction between group ability distribution and item discrimination. All procedures performed well with unequal ability distributions and low and moderate item discrimination, but poorly with high studied item discrimination; SIBTEST showed the best Type I Error control under these conditions. Similarly, for the descriptive procedures, the $STND_{ES\text{-}DIF}$ performed best when group ability distributions were equal, but when groups had unequal ability distributions, the modified SIBTEST DIF effect size measure performed best.

Welch and Hoover (1993) compared the performance of the Mantel to HW1 and HW3. The matching test consisted of multiple choice items and one studied item, and the level of

uniform DIF in this item ranged from 0 to 0.45. They also studied the effect of unequal ability distributions and sample size and report that under conditions of equal ability, HW1 and HW3 had higher power than the Mantel procedure but that Type I Error rates were high when ability distributions were unequal. The Mantel had good Type I Error control (Welch & Hoover, 1993).

Spray and Miller (1994) simulated a 20 item test to compare the power and Type I Error of LDFA, Mantel, and GMH for detecting DIF in polytomous items in small (n = 500) and large (n = 2000) samples. All twenty items were polytomous; the last item was the studied item. Uniform DIF, balanced DIF, and nonuniform DIF were simulated to test power. The Type I Error rate for the DIF detection procedures was defined as the number of times out of 100 that a significant test was observed for Items 1 - 19 (non-DIF); the rate for all three procedures was low under these study conditions. LDFA had higher power than the other two methods for uniform and non-uniform DIF; it also had higher power for balanced DIF in large samples. However, the GMH had higher power for detecting balanced DIF in small samples.

Chang, Mazzeo, and Roussos (1996) conducted two studies to compare the polytomous SIBTEST procedure to the Mantel and inferential STND procedures. The first study was a replication of that done by Zwick, Donoghue, and Grima (1993); the same 54 conditions were employed, and 600 replications were performed for each condition. In Study One (Chang et al, 1996), differences between the three procedures were small; however, the Mantel and $STND_{ES\text{-}DIF}$ had better Type I Error control and slightly higher power for uniform DIF than Poly-SIBTEST. Type I Error of the Mantel averaged 0.049 across all conditions, including group ability differences, Type I Error rates for the $STND_{ES\text{-}DIF}$ averaged 0.046, and Type I Error for SIBTEST

averaged 0.063. None of the procedures had adequate power to detect balanced DIF, low shift, or high shift DIF.

In the second study, the effect of differing item discrimination levels on the performance of the three procedures under conditions of group ability differences (Reference = $N(0,1)$, Focal = $N(-1, 1)$ was investigated. Eleven values for studied item discrimination, ranging from 0.15 to 2.00, were used. In total, 22 different conditions were studied, and 1000 replications were employed for each condition. In this second study, the Mantel and $STND_{ES\text{-}DIF}$ had unacceptably high Type I Error rates when studied item discrimination differed from the average for the matching test. Poly-SIBTEST was more robust under these conditions.

French and Miller (1996) conducted a simulation study to compare three different methods of coding for polytomous logistic regression: continuation ratio logits, cumulative logits, and adjacent categories logits. In this study, several versions of a simulated polytomous 25-item test were generated; the 25th item was the studied item. Three different levels of nonuniform DIF were simulated; differences in *a* parameters were 0.5, 1.0, and 1.5. One level of balanced DIF was simulated. Type I Error was not studied. French and Miller (1996) found that the adjacent categories coding scheme had very little power for detecting DIF but that the continuation ratio and cumulative logits coding schemes had good power under most conditions. Power increased with sample size for all three methods. French and Miller (1996) concluded that the adjacent categories coding scheme lost too much information and was not powerful enough to use. They also concluded that although the other coding methods had high power, the necessity of performing separate regressions for every score threshold made logistic regression unwieldy and difficult to interpret.

Tian (1999) performed a large simulation study (432 cells) in which she compared the power and Type I Error of Mantel, GMH, the inferential test for $STND_{ES\text{-}DIF}$ derived under the hypergeometric model, Poly-SIBTEST, and LDFA. She generated 20 and 40 item tests; four-fifths of the items on each test were dichotomous and the others were polytomous with four score levels. One of the polytomous items was the studied item. The following conditions were studied; sample size (600, 1800, and 2400) and sample size ratio (1:1 or 2:1), difference in group ability distributions (0, 0.5, and 1.0), studied item discrimination (0.5, 1.0, and 1.5), and type of DIF (uniform, nonuniform, or balanced). In the uniform DIF condition, the difference between *b*-parameters was 0.25. The difference in item discrimination in the nonuniform DIF condition ranged from 0.5 to 1.5. Tian (1999) found that group ability difference was the major factor influencing the performance of all procedures. As group ability differences increased, Type I Error and power both increased; this effect was magnified by high item discrimination. LDFA-nonuniform seemed most susceptible to group ability differences; it had unacceptably high Type I Error under all conditions with moderate and large differences in ability distributions. Poly-SIBTEST had somewhat better Type I Error control under these conditions than the other procedures. When group abilities were equal, all procedures had good Type I Error control; however, poly-SIBTEST had somewhat higher Type I Error rates than the other procedures, especially with the short test. The STND was, however, affected by sample size ratio and had lower Type I Error when the ratio was higher.

Due to the overriding effect of group ability distribution, power was compared only under conditions of equal ability distribution. All procedures had high power for detecting uniform DIF; LDFA-uniform had slightly higher power than the other procedures. The Mantel, STND,

SIBTEST, and LDFA-uniform procedures were completely unable to detect balanced DIF. The GMH and LDFA-nonuniform were both able to detect balanced DIF, but the GMH was more powerful. However, neither GMH nor LDFA-nonuniform worked well with small sample size or low item discrimination. Mantel, $STND_{ES\text{-}DIF}$, SIBTEST, and LDFA had no power to detect nonuniform DIF. Both the LDFA-nonuniform and GMH were able to detect nonuniform DIF; LDFA nonuniform was more powerful than GMH when there was no difference in group ability. However, GMH had higher power when group ability differences existed; the power of the LDFA deteriorated as group ability differences increased. Tian (1999) concluded that, overall, the LDFA procedures were best.

Ankenmann, Witt, and Dunbar (1999) conducted a simulation study to compare the Type I Error and power for uniform and balanced DIF of the Mantel and IRT-Likelihood Ratio (IRT-LR) procedures. Three factors were manipulated: sample size (total 1000, 2500, and 4000), differences in group ability distribution (0 and 1.0 standard deviations), and item discrimination. Several 26-item tests, with a mixture of polytomous and dichotomous items were generated; the 26^{th} item was always the studied item. Six studied items were generated by crossing two discrimination parameter values with different levels of DIF: null DIF, uniform DIF, and balanced DIF. The difference in *b*-parameter values between the reference and focal groups was 0.25 Three combinations of sample size and sample size ratios were investigated: 2000/2000, 2000/500, and 500/500. In total, 72 different conditions were simulated; 100 replications were done for each condition. Both Mantel and the IRT-LR had low Type I Error rates when group ability distributions were identical and the studied item was included in the matching test. However, when ability distributions were unequal, the Type I Error of the Mantel exceeded the

nominal rate while those for the IRT-LR procedure remained close to nominal values. Both procedures had high power for uniform DIF; the Mantel had higher power when item discrimination was moderate. The Mantel could not detect balanced DIF, while the IRT-LR procedure had moderate to good power. The IRT-LR had inadequate power with small sample sizes.

Summary of Empirical Findings

Nine studies on polytomous DIF detection procedures were found; the magnitude of DIF simulated in them is summarized in Table 3.

Table 3. The Magnitude of Polytomous DIF in Previous Monte Carlo Studies

Study	Difference in *b* parameters	Difference in *a* parameters
Zwick et al, 1993 American Physics Board Exam	0.1 and 0.25	None
Zwick et al, 1997	0.25	
Welch & Hoover, 1993		
Spray and Miller, 1994	0.25	0.5
Chang et al, 1996	0.10 and 0.25	None
French and Miller, 1996	1	0.5, 1.0, and 1.5
Tian, 1999	0.25	0.5, 1.0, and 1.5
Ankenmann et al, 1999	0.25	

The Mantel

The Mantel has been evaluated in seven studies (Ankenmann et al., 1999; Chang et al., 1996; Spray & Miller, 1994; Tian, 1999; Welch & Hoover, 1993; Zwick et al., 1993; Zwick et al,

1997). Type I Error for the Mantel procedure has been low and acceptable (< 0.1) under most conditions (Chang et al., 1996; Spray & Miller, 1994; Tian, 1999; Zwick et al., 1993). The performance of the Mantel does not seem to be affected by moderate (- 0.5 standard deviation) differences in ability distributions (Tian, 1999), and some researchers (Ankenmann et al, 1999: Chang et al., 1996; Zwick et al., 1993) report only small increases in Type I Error with large group ability differences. However, Tian (1999) reported high Type I Error rates when group ability differences were large. Studied item discrimination may also affect Type I Error; Chang et al. (1996) report high Type I Error when the studied item discrimination differed from the mean of the nonstudied items. There is an interaction between large group ability differences and high studied item discrimination; when group abilities are unequal and item discrimination is high, Type I Errors are increased; values typically range from around 0.20 to 0.55.

The Mantel has high power for detecting uniform DIF under most conditions. Zwick et al. (1993) found that the Mantel had a higher DIF detection rate than the GMH for uniform DIF when DIF was of low (0.10) and moderate (0.25) magnitude. Chang et al. (1996) found that the Mantel had a somewhat higher detection rate for uniform DIF than SIBTEST and STND under the same conditions. Tian (1999) and Spray and Miller (1994) found that the Mantel had higher power than the GMH for detecting uniform DIF; in Tian's study power was comparable to that of SIBTEST and STND. However, the Mantel had slightly lower power than LDFA for detecting uniform DIF (Spray & Miller, 1994; Tian, 1999).

The Mantel was not designed to detect nonuniform DIF, and researchers have found that it has no power (0.00 to 0.05) to do so (Spray & Miller, 1994; Tian, 1999). It is not useful for the

detection of balanced DIF either (Chang et al., 1996; Spray & Miller, 1994; Tian, 1999; Zwick et al., 1993).

STND

The $STND_{ES\text{-}DIF}$ procedure has been examined in four studies (Chang et al, 1996; Tian, 1999; Zwick et al, 1993; Zwick et al, 1997); the descriptive statistic was studied in two of these, and the inferential statistic was studied in three. Reported values for Type I Error are somewhat lower than those for the Mantel; power values are nearly identical. Like the Mantel, $STND_{ES\text{-}DIF}$ has high power for detecting uniform DIF under most conditions, except when ability distributions are unequal and studied item discrimination differs from the mean of the matching items. The $STND_{ES\text{-}DIF}$ procedure was not designed to detect nonuniform DIF; it has no power to detect either non-uniform DIF or balanced DIF.

The GMH

The GMH procedure has been evaluated in three studies (Spray & Miller, 1994; Tian, 1999; Zwick et al, 1993). The Type I Error rate of GMH is generally very good under most conditions. For example, Tian found that with moderate differences in group ability, the GMH had lower Type I Error rates than the Mantel, $STND_{ES\text{-}DIF}$, and LDFA (uniform and nonuniform). However, as with other procedures, Type I Errors were unacceptably high when ability distributions were highly unequal and/or item discrimination was high (Tian, 1999).

In general, the GMH has somewhat lower power than the Mantel, $STND_{ES\text{-}DIF}$, and LDFA for detecting uniform DIF. The GMH has shown higher power for detecting balanced DIF than most of the other procedures; power ranged from 0.05 to 0.49 when sample size and item

discrimination were low (Tian, 1999, Zwick et al., 1993). The GMH also had excellent power for detecting balanced DIF with large sample sizes (Spray & Miller, 1994).

The GMH has higher power than the Mantel (Spray & Miller, 1994; Tian, 1999), STND, (Tian,1999) and SIBTEST procedures (Tian, 1999) for detecting nonuniform DIF under all conditions.

HW1 and HW3

Little empirical evidence is available on the performance of HW1 and HW3. Welch and Hoover (1993) report that these statistics had higher power for uniform DIF than the Mantel, but that they also had high Type I Error rates under conditions of unequal abilities.

LDFA

LDFA has only been evaluated in two Monte Carlo studies, but seems promising. Type I Error control is generally good under most conditions. However, Type I error may be problematic when group ability differences are present; under these conditions, the Type I Error rate of the LDFA-nonuniform statistic was higher than that of other procedures (Tian, 1999).

In small samples, LDFA-uniform has slightly higher power than other procedures for detecting uniform DIF; in large samples, it does as well or better than other procedures (Spray & Miller, 1994; Tian, 1999).

LDFA-nonuniform has good power for detecting nonuniform DIF under most conditions. Miller and Spray (1993) found that power ranged from 0.44 in small samples to 1.00 in large samples; this was better than GMH. Tian (1999) also found that LDFA-nonuniform had high power for detecting non-uniform DIF; power ranged from 0.28 to 0.99, depending on sample size, item discrimination, and group ability difference. LDFA-nonuniform had unacceptably low

power to to detect nonuniform DIF when there were differences in ability between the reference and focal groups. For example, the mean power of LDFA-nonuniform to detect a difference of 1.5 in *a*-parameters was 0.84 with no group ability difference. This decreased to 0.66 when there was a 0.5 difference in group ability, and 0.46 with a large difference in group ability distribution.

LDFA-nonuniform has moderately high power to detect balanced DIF except in small samples (Spray & Miller, 1994; Tian, 1999).

SIBTEST

Polytomous SIBTEST has been evaluated in three studies. When ability distributions are equal, or only moderately different (- 0.5 standard deviations), Type I Error is comparable to the Mantel, GMH, and LDFA procedures. Poly-SIBTEST works well for detecting uniform DIF, but in general, has slightly lower power than other procedures. However, SIBTEST is generally more robust in terms of Type I error than other procedures when ability distributions are unequal (Zwick et al., 1993).

Poly-SIBTEST was not designed to detect nonuniform DIF, and has no power to do so (Chang et al., 1996; Tian, 1999). Its power for detecting balanced DIF is also quite low (Chang et al., 1996; Tian, 1999).

Logistic Regression

Polytomous logistic regression has only been evaluated in one study by French and Miller who evaluated the coding schemes proposed by Agresti (1990), but their study was limited. For example, Type I Error was not considered. Ordinal Logistic Regression (Zumbo, 1999) has not been evaluated in a Monte Carlo study.

Summary of Factors Affecting DIF Detection

Sample Size

In general, as sample size increases, so does power. Type I Error is less affected by sample size than is power, but does increase with sample size when there are differences in group ability distributions (Tian, 1999).

The effect of sample size ratio (between reference and focal group) has not been well studied. Tian (1999) found that a ratio of 2:1 generally had little effect on DIF detection. However, unequal sample sizes did result in decreased power for some procedures.

DIF Magnitude

Power increases markedly as the magnitude of DIF increases. For example, Zwick, Donoghue, and Grima (1993) and Chang, Mazzeo, and Roussos (1996) found that power for uniform DIF increased from below 0.20 to at least 0.70 when the difference between *b*-parameters increased from 0.1 to 0.25.

Differences in Group Ability Distributions

Moderate differences in group ability distribution have little effect on Type I Error rates. Results are somewhat conflicting for large (1.0) differences. Both Chang et al. (1996) and Zwick et al. (1993) found that large group ability differences had very little effect on Type I Error, unless studied item discrimination differed from the mean. However, Tian (1999) reports a strong main effect of group ability difference; Type I Errors increased markedly when group ability differences were large. Group ability differences have little effect on power for uniform DIF. However, the power of the LDFA- nonuniform statistic to detect nonuniform DIF decreased as

between-group ability differences increased. The effect of group ability distribution is magnified when studied item discrimination is high.

Studied Item Discrimination

Studied item discrimination may affect DIF detection in several ways. Type I Error is increased when studied item discrimination is markedly different from that of the mean for the matching test (Chang et al., 1996; Zwick et al., 1997). Power for uniform DIF is low when studied item discrimination is low, but very high when studied item discrimination is high (Chang et al., 1996). There is also a strong interaction between studied item discrimination and group ability differences; this interaction results in high Type I Error when studied item discrimination is high and when there are group ability differences (Chang et al., 1996; Zwick et al., 1997).

Skewness. The effect of skewness on DIF detection has only been evaluated in one study. Monaco (1997) compared Type I Error and power of Lord's Chi-Square (Lord, 1980), Mantel-Haenszel, and an IRT-based procedure called DFIT (Raju, Van der Linden, & Fleer, 1996) for detecting DIF in dichotomous items. She also studied the effect of moderate skewness (Chi-square with 15 degrees of freedom) and high skewness (Chi-square with 5 degrees of freedom) on DIF detection rates. She found that Lord's Chi-Square, the Mantel-Heanszel, and DFIT all had high power to detect uniform DIF, but that none had high power to detect nonuniform DIF in dichotomous items. Moderate skewness had little effect on DIF detection and Type I Error while high skewness resulted in a 5 to 10% decrease in power.

Summary

The Mantel, $STND_{ES\text{-}DIF}$, HW1 and HW2, SIBTEST, GMH, LDFA, and Ordinal Logistic Regression procedures are all straightforward to apply, although GMH may be difficult to interpret. The Mantel, Poly-SIBTEST, and $STND_{ES\text{-}DIF}$ do not detect either nonuniform or balanced DIF. The GMH works nearly as well for uniform DIF as Mantel and STND; it works fairly well for balanced DIF and can detect nonuniform DIF better than all other procedures except LDFA-nonuniform. Poly-SIBTEST only detects uniform DIF, but it may have better Type I Error control when group abilities are different. LDFA-uniform and LDFA-nonuniform seem promising, although LDFA-nonuniform has poor Type I Error control when there are group ability differences. Empirical work on polytomous logistic regression has been limited, although French and Miller (1994) did find that cumulative logits and continuation ratio coding had high power for both uniform and nonuniform DIF.

Gaps in Existing Research

Although a number of techniques for assessing DIF in polytomous items exist, with the exception of the Mantel, they have not been well studied. Most have only been tested in one or two studies, and polytomous logistic regression has only been evaluated in one study. Ordinal Logistic Regression (Zumbo, 1999) seems promising, but has not yet been evaluated in a Monte Carlo study. Thus, its effectiveness is unknown. More empirical work is needed in order to determine which procedures are best for a given set of study conditions (Potenza & Dorans, 1995). Also, nonuniform DIF has not been well studied; only French and Miller (1996), Spray

and Miller (1993) and Tian (1999) have studied the effectiveness of various procedures for detecting nonuniform DIF.

Conditions which have been studied thus far are quite limited. For example, although highly uneven reference and focal group sample sizes may exist in many situations, only one researcher (Tian, 1999) has evaluated the effect of unequal group size, and she only considered a 2:1 ratio. Study of the effect of different sample sizes, differences in group ability distributions, skewed ability distributions, and studied item discrimination is needed to more fully understand how these procedures might perform in applied testing situations.

The purpose of the present study is to add to the body of knowledge on DIF assessment procedures for polytomous items by using simulated data reflective of realistic testing situations to compare the Type I Error and power of four procedures: the Mantel, the GMH, LDFA (uniform and nonuniform), and Ordinal Logistic Regression (total, uniform, and nonuniform).

Research Questions

Four main research questions and two secondary questions were addressed in this study:

1. What are the Type I Error rates for the Mantel, GMH, LDFA, and Ordinal Logistic Regression when they are used to detect DIF in polytomous item responses simulated to have no DIF?

It was hypothesized that the procedures would all have, on average, very low Type I error rates (Hypothesis 1.1), but that the OLR would have relatively the lowest Type I error rates (Hypothesis 1.2). This hypothesis was based on previous research, and also on the fact that the way in which DIF is modelled in OLR closely matches the theoretical definition of DIF.

2. How powerful are the Mantel, GMH, LDFA, and Ordinal Logistic Regression procedures when they are used to detect DIF in polytomous item responses simulated to have uniform and nonuniform DIF?

It was expected that the Mantel would have the highest power for uniform DIF (Hypothesis 2.1), but that it would not be useful for detecting nonuniform DIF (Hypothesis 2.2). The GMH was expected to be very powerful in detecting both uniform and nonuniform DIF (Hypothesis 2.3). It was also hypothesized that LDFA and Ordinal Logistic Regression could detect both uniform and nonuniform DIF (Hypothesis 2.4) and that these procedures would have the highest power for nonuniform DIF (Hypothesis 2.5). These hypotheses were largely based on previous research; hypotheses about the OLR were based on the fact that the OLR is closely related to both the GMH and LDFA.

3. *Do group ability distribution differences and skewness, sample size ratio, and item discrimination affect the Type I Error and power of the DIF assessment procedures?*

Based on previous research, it was hypothesized that differences in group ability distributions would result in increased Type I Error when item discrimination was high (Hypothesis 3.1), and that differences in sample size would lower power (Hypothesis 3.2).

4. *What effect does the inclusion of effect size in the Mantel, GMH, LDFA, and OLR have on Type I Error? On power to detect uniform and nonuniform DIF?*

For all procedures, it was hypothesized that including effect size in the decision rule would result in substantially lower Type I error for all procedures (Hypothesis 4.1), as well as slightly lower power for uniform and nonuniform DIF (Hypothesis 4.2). This hypothesis was based on

the fact that the requirement of an important difference between groups will result in fewer items being labelled as DIF.

Two secondary questions concerned the second phase of DIF detection: classification of DIF as predominantly uniform or nonuniform DIF. Because the Mantel and GMH only have one test, the questions about classification were not relevant for them.

5. What are the Type I Error rates for the LDFA and OLR uniform tests in data simulated to have only nonuniform DIF (i.e. how often is nonuniform DIF misidentified as uniform DIF)

6. What are the Type I Error rates for the LDFA and OLR nonuniform tests in data simulated to have only uniform DIF (i.e. how often is uniform DIF misidentified as nonuniform DIF?)

Rationale for the Selection of the DIF Assessment Procedures under Study

The Mantel was included as a standard because it is the most common DIF assessment technique, and has high power for uniform DIF. It should provide a basis for comparison with other studies. The GMH was included because it has high power for uniform and nonuniform DIF. The LDFA was studied because it has been shown to have high power for nonuniform DIF as well as for uniform DIF, and because it should be useful in discriminating between different types of DIF. The OLR was included because it is a practical procedure which should be useful for the detection of both uniform and nonuniform DIF and for differentiating between uniform and nonuniform DIF. Because its validity as a DIF detection procedure has not been previously assessed, this study will provide new information on DIF detection.

Except for the Mantel, the methods chosen are among the most promising and practical for the detection of both uniform and nonuniform DIF. All of the techniques under study are easily

transferred to applied situations as they are readily available and easily understood and applied. Furthermore, they do not require the purchase of special computer programs.

Poly-SIBTEST could not be studied because source FORTRAN code was unavailable; thus the software could not be linked to the other programs which were used in the study. Moreover, the polytomous procedure is only useful for detecting uniform DIF, and the program is costly. IRT based procedures are not included because they are difficult and costly to use in most applied settings.

The rationale for studying effect size is that in hypothesis testing, large sample sizes can lead to a statistically significant test in the absence of a meaningful effect. Because of this, test statistics should be accompanied by measures of effect size which show the strength of association between the dependent and independent variables (Kirk, 1996; Zumbo, 1999). The evaluation of DIF procedures for polytomous items has largely been done on inferential tests; less work has been done on measures of effect magnitude (Zwick et al., 1997). This study will provide one of the first evaluations of the impact of including effect size in DIF decision rules.

Comparison of Studied Techniques

There are methodological similarities among the techniques; there are also many differences. All use total test score rather than latent ability as the matching variable. In all, the studied item must be included in the matching test. LDFA and OLR are closely related; both are based on logistic regression, and both model uniform as well as nonuniform DIF. The main difference between the two is that with OLR, item score is the dependent variable, and total score, group membership, and group * total are predictors whereas in LDFA, group membership is the

dependent variable. LDFA is the only DIF assessment procedure to model group membership rather than item score. The Mantel is also related to LDFA and OLR. It can be seen as based on a logistic regression model where the ability variable is discrete and no interaction term is allowed (Swaminathan & Rogers, 1990). The Mantel and GMH differ from each other in that the Mantel is based on a comparison of group means while the GMH takes the total range of scores into account. This means that the Mantel only detects uniform DIF and the GMH can be used to detect uniform, nonuniform, and balanced DIF.

CHAPTER III
METHODS

Overview

The methods used to address the research questions are presented in this chapter, which is divided into five sections. In the first section, the procedures under study are presented, and details of their method of application are given. The second section describes independent variables, conditions which were held constant, and an overall summary of the study design. In the third section, steps in data generation and testing of the DIF assessment programs are outlined. In the fourth section, methods and results from validation of the data generation and DIF procedure programs are presented, while in the fifth section details of the final analyses are presented.

The OLR, LDFA, Mantel, and the GMH were studied and compared. In each of these procedures, total test score was used as the matching variable; the studied item was included in that total. Refinements were made to the LDFA and OLR procedures. The Mantel and GMH were applied in the standard manner, except that a measure of effect size was added to ensure comparability and to study the impact of effect size on each procedure. All measures were evaluated with and without effect size. Full details are below.

Modelling in LDFA and OLR

Swaminathan and Rogers (1990) and Zumbo (1999) have focused on the two-degree of freedom test of the difference between the fully saturated model and the null model as the logical first step in DIF assessment. If DIF is identified by this overall test, then tests of the difference between the nonuniform and uniform models are used to classify the type of DIF. This sequential

modelling strategy is both conceptually sound and intuitively reasonable. It makes good sense to first determine whether or not there is DIF in the item and then to classify it. Therefore, as suggested by Zumbo (1999), a sequential modelling strategy was followed for both the OLR and LDFA. This involved two main steps. The first step was identification. In this step, an overall, two-degree of freedom χ^2 test of the difference between the nonuniform model and the null model was used in conjunction with an effect size measure to identify DIF of any kind. The second step was classification of DIF as uniform or nonuniform; this was only done if there was DIF in the item.. Then, the-degree of freedom χ^2 test of the difference between the uniform and the null model was performed (with effect size). If this was significant, the DIF was said to be predominantly uniform. If it was not, the DIF was said to be nonuniform. Thus, in this study, three tests were evaluated for both the OLR and LDFA: the overall two-degree of freedom test of the difference between nonuniform and null models, a test of uniform DIF (the difference between uniform and null models), and a test of nonuniform DIF (significant total DIF, but nonsignificant uniform DIF). These are described below.

Ordinal Logistic Regression

Two modifications were made to the procedure which was developed by Zumbo (1999) and presented in the literature review for this study. First, the measure of effect size was changed from R^2 to one based on weighted group differences in item score due to DIF; separate effect size measures were developed for uniform and nonuniform DIF. Second, the cumulative logits were each tested in separate regressions rather than in one overall regression.

The Effect Size Measures

While R^2 is frequently used and is easily understood by researchers, it is not directly interpretable in terms of the amount of DIF in an item. Therefore, two alternate effect size

measures, one for uniform DIF and one for nonuniform DIF, were developed by Aylesworth and Kristjansson (2000). These measures both give an indication of the difference in item score between the reference and focal groups, controlling for total score. The measure for total DIF is given by:

$$\text{Total group difference in item score} = \frac{1}{N}\left(\sum_{m}^{I} N_m |D_m|\right) \tag{36}$$

Where:

N = total sample size

N_m = the number of examinees at each total score level and

D_m = the total difference in regression predicted score values between the reference and focal group at each level of the total score.

Because the absolute value for the difference between reference and focal groups at each level is used, the total includes all differences, no matter in what direction. Thus, both uniform and nonuniform DIF should be detected with this effect size measure.

The measure for uniform DIF is given by:

$$\text{Mean difference in item score} = \left|\frac{1}{N}\left(\sum_{m}^{I} N_m D_m\right)\right| \tag{37}$$

where:

N = total sample size

N_m = the number of examinees at each total score level.

D_m = mean difference in regression predicted score values between the reference and focal group at each level of the total score.

It can be seen from this formula that if some differences in predicted score values are negative, and some are positive, they will cancel each other out. Therefore, this effect size measure will only detect uniform DIF.

This measure for uniform DIF is similar to $STND_{ES\text{-}DIF}$ in that it reflects differences in item score due to DIF. However, it differs from $STND_{ES\text{-}DIF}$ in three ways: 1) the estimated score values are predicted from a regression model while in $STND_{ES\text{-}DIF}$ they are predicted from observed scores 2) in our effect size measure, difference scores are weighted by the total number of subjects at a given score level while in $STND_{ES\text{-}DIF}$ difference scores are weighted by focal group frequencies and 3) this measure is parametric while the $STND_{ES\text{-}DIF}$ is non-parametric.

These formulae have several advantages. First, being based on differences in item scores, they are easily understood and interpreted and show the amount of difference that DIF will make to the item score and to the total test score. Second, they will enable psychometricians to base identification of DIF on meaningful differences in item score as well as on significance; the cut-off value can be chosen to be relevant for a given setting. Finally, these measures are easily derived from procedures in SAS and SPSS.

Testing Cumulative Logits Separately

In the OLR procedure outlined by Zumbo (1999), each regression provides a simultaneous test of all cumulative logits (one for each threshold); comparison of the χ^2 and effect size measure from these overall equations shows whether there is any DIF present, and then, whether the DIF is predominantly uniform or nonuniform. However, the simultaneous test of cumulative logits can only be performed if the assumption of equal slopes is upheld. Preliminary testing revealed that

the assumption of equal slopes was violated in 75% of the datasets generated for this study; this violation resulted in a substantial loss of power. Therefore, it was necessary to modify the OLR procedure so that each cumulative logit would be tested separately. In order to circumvent the problems with interpretation described by French and Miller (1996), the Bonferonni correction was applied; an item was said to have significant DIF if any one of the three logits was significant at the 0.05/3 level and if the effect size was above 0.03. These rules ensured that the decision about DIF was straightforward, even though several individual regressions were run.

The OLR procedure

The OLR procedure used in this study involved the following steps:

Step One: OLR-overall. The significance of the difference between the fully saturated model and the null model was tested. This involved comparing an equation which modeled null DIF for each logit to one which specified nonuniform DIF (total score group + total * group) for each logit. If any of the three differences in logits between the nonuniform models and the null models was significant at the 0.05/3 level, the effect size measure for total DIF (Equation 36) was calculated. The specified cut-off was 0.03. Thus, if there was statistically significant DIF, and if the total difference in item score between the reference and focal groups, controlling for total score, exceeded 0.03 of a point (out of 3), the item had DIF.

Step Two: OLR-uniform. When DIF was flagged in the studied item, testing for uniform DIF followed. This test was performed by fitting an equation which specified uniform DIF only (total score + group) for each logit and comparing it to an equation which specified null DIF (total score only) for each logit. If any of the three differences in logits

were significant at the 0.05/3 level, and if the effect size for uniform DIF was above the cut-off of 0.03, the item was said to display uniform DIF.

Step Three: OLR-nonuniform. If the item had overall DIF, but the criteria for uniform DIF were not met, the item was said to display nonuniform DIF.

A second series of analyses was used to examine the performance of the OLR procedure in the absence of the effect size measure. In these analyses, the same steps were followed, except that statistical significance alone was the basis for decision making.

Logistic Discriminant Function Analysis

Two modifications were made to the LDFA. First, measures of effect size were developed and used in the decision rules for LDFA. Second, cubic and quadratic terms were added to enhance the fit of the model.

Effect size measures for LDFA

These measures are based on the mean weighted probability of group membership at each level of total score. The measure for total DIF is:

$$\mu \text{ (difference in probability of group membership)} = \frac{1}{N}\left(\sum_{i}^{1} N_i |D_i|\right) \qquad (38)$$

where:

N_i = the number of subjects with total score = i

N = the sum of all N_i

$D_i = p_{fi} - p_{ri}$

p_{fi} = probability of membership in focal group when total score = i and

p_{ri} = probability of membership in reference group when total score = i

The measure for uniform DIF is given by:

$$\mu \text{ (difference in probability of group membership)} = \left| \frac{1}{N} \left(\sum_{i}^{1} N_i D_i \right) \right| \qquad (39)$$

where:

N_i = the number of subjects with total score = i

N = the sum of all N_i

$D_i = p_{fi} - p_{ri}$

p_{fi} = probability of membership in focal group when total score = i and

p_{ri} = probability of membership in reference group when total score = i

$D_i = p_{fi} - p_{ri}$

Adding cubic and quadratic terms

Preliminary testing revealed that the LDFA procedure had an extremely high Type I Error rate (ranging from 0.40 to 0.90) when ability distributions were different and skewness was present in the data. Therefore, cubic and quadratic terms for the total score were added to the null, uniform, and nonuniform models in order to improve fit of the null model.

The LDFA procedure

The following measures were tested:

Step One: LDFA-overall. This two-degree of freedom χ^2 test statistic is a test of the difference between Equations 23 and 25 (difference between the nonuniform and null models). This value had to exceed 5.99 to reach significance. If DIF was significant, and if the difference

in overall probability of group membership exceeded 0.16, the item was said to have DIF. The cut-off of 0.16 was based on preliminary testing.

Step Two: LDFA-uniform. This one-degree of freedom χ^2 test statistic is a test of the difference between Equations 24 and 25 (difference between the uniform and null model). The value had to exceed 3.84 to reach significance. The item was said to have uniform DIF if the test for uniform DIF was significant, and if the effect size for uniform DIF exceeded 0.16.

Step Three: LDFA-nonuniform. The item was said to have nonuniform DIF if the criteria for total DIF were met, and if the criteria for uniform DIF were not met.

In a second series of analyses the performance of the LDFA procedure in the absence of the effect size measure was examined. In these analyses, the same steps were followed, except that statistical significance alone was the basis for decision making.

The Mantel

Calculation of the Mantel was based on a comparison of the item means for the two groups at each level of total score (Mantel, 1963). The null hypothesis is that at a fixed level of total score, there is no conditional association between the item score and group membership. The test statistic for Mantel is distributed as a χ^2 with one-degree of freedom. The Aylesworth and Kristjansson (2000) effect size measure for uniform DIF (Equation 37) was used with the Mantel. In the first set of analyses, the Type I error and power for uniform and nonuniform DIF of the Mantel with effect size were examined. In these analyses, the χ^2 value had to exceed 3.84 and the mean difference in predicted item score had to exceed 0.03 for the item to have DIF. A second series of analyses was used to examine the performance of the Mantel procedure in the

absence of the effect size measure. In these analyses, the same steps were followed, except that statistical significance alone was the basis for decision making.

The GMH

The null hypothesis for the GMH is that for a given level of test score, there is no association between item responses and group membership (Potenza & Dorans, 1995). The GMH statistic is distributed as a χ^2, with k-1 degrees of freedom. The Aylesworth and Kristjansson effect size for total DIF (Equation 36) was added as a measure of effect size. In these analyses, the χ^2 had to exceed 7.81 and the effect size measure for total DIF had to exceed 0.03 for the item to have DIF.

In a second series of analyses, the performance of the GMH procedure in the absence of the effect size measure was examined. In these analyses, the same steps were followed, except that statistical significance alone was the basis for decision making.

Study Design

The Mantel, GMH, LDFA, and OLR procedures were tested using a simulated 26 item test; all items were polytomous. The 26th item was the studied item. Examinees were matched on total test score; this 'matching test' always included the first 25 items as well as the studied item.

Independent Variables

Several factors were varied in order to assess their effect on the performance of the DIF detection procedures. These included: the presence and type of DIF in the studied item, studied item discrimination, reference and focal group sample size ratio, difference in group ability distribution, and skewness of the ability distributions. Other factors, including the number of test

items, the number of response categories in items, and parameters for the nonstudied items were held constant.

Presence and Type of DIF

Three DIF conditions were simulated in the studied item: no DIF (null DIF), uniform DIF, and nonuniform DIF. Items with no DIF were generated to evaluate the Type I Error of each procedure. Items with DIF were generated to test the power of each procedure for detecting uniform DIF, and nonuniform DIF, and for differentiating between uniform and nonuniform DIF. After careful consideration of how DIF may occur and consultation with experts (Spray, personal communication, 1999; Zumbo, personal communication, 2000), balanced DIF was not simulated because it is highly unlikely that it would be found in real data.

In the null DIF condition, the *a* and *b* parameters were equivalent for the reference and focal groups. In the uniform DIF condition, *a*-parameters for both groups remained the same, but *b*-parameters were increased by 0.25 at each score level for the focal group; this meant that, at each transition, the focal group was less likely than the reference group to achieve the higher score (1 versus 0). This magnitude of DIF was based on levels used in other simulation studies (see Table 2) and on that found in real data (see Table 1). The 0.25 difference in *b*-parameters resulted in a mean difference between groups of 0.12 (out of a possible 3 points) in the total item score. In the nonuniform DIF condition, *b*-parameters remained the same, but the *a*-parameter for the reference group was increased by 1.0 when studied item discrimination equaled 0.8, by 1.3 when studied item discrimination equaled 1.2, and by 1.6 when studied item discrimination was 1.6. These values were chosen to make the DIF magnitude in the uniform and nonuniform conditions approximately equivalent. Swaminathan and Rogers (1990) used a similar approach to produce equivalent levels of uniform and nonuniform DIF in simulated data.

Studied Item Discrimination

Item discrimination may have a significant impact on Type I Error and on power, particularly when reference and focal groups differ in ability. In this study, three item discrimination levels were evaluated; 0.8, 1.2, and 1.6. These represent typical low, moderate, and high item discriminations.

Group Sample Size Ratio

The power and Type I Error of many inferential statistics may be affected by differences in group sample size (Checkoway, Pearce, & Crawford-Brown, 1989). However, the effect of group sample size ratio on DIF detection has not yet been well studied. Tian (1999) found that the power of the STND to detect uniform DIF was decreased slightly when sample sizes were unequal.

Two different sample size ratios were used in the present study. In the equal condition, both the reference and focal groups had sample sizes of 2000. In the unequal (4:1 ratio) condition, the reference group had 3200 subjects and the focal group had only 800.

Differences in Group Ability Distributions

Although results are inconsistent, it seems that large differences in mean ability between the reference and focal group may result in higher Type I Error (Tian,1999). There is also an interaction between group ability difference and studied item discrimination. In the present study, two levels of difference in mean ability (total test score standardized) were studied: 1) Equal reference and focal group ability distributions, where the reference and focal group had standard normal ability distributions with a mean of 0 and a standard deviation of 1.0) and 2) Unequal distributions where the reference group had a standard normal distribution of ability (mean 0 and

standard deviation of 1, and the focal group had a mean ability of - 0.5 and a standard deviation of 1.

Skewness in Ability Distributions

Ability distributions are often skewed in real item response data (Nandakumar & Yu, 1996). Nonetheless, the impact of skewness on the power and Type I Error of polytomous DIF detection methods has only been evaluated in one other study (Monaco, 1997). In the present study, two levels of skewness were studied: a moderate negative skew (- 0.75) in ability distributions for both the focal and reference groups and 2) no skewness. The total number of conditions = 96 (12 studied item levels x 2 sample size ratios x 2 skewness levels x 2 ability differences). A summary of these factors is shown in Table 4.

Table 4. A Summary of the Study Design, including all factors to be varied

Studied item: 3 DIF Conditions x 3 Studied Item Discrimination Levels (9 total)
1. Null DIF: 3 items (one for each discrimination level: $a_{jF} = a_{jR} = 0.8, 1.2,$ or 1.6) $b_{jkF} = b_{jkR} = (-1.5, 0,$ and $1.5, k = 1,2,3,4)$ 2. Uniform DIF (constant): 3 items ($a_{jF} = a_{jR} = 0.8, 1.2,$ or 1.6, $b_{jkF} = b_{jkR} + 0.25$) 3. Nonuniform DIF: 3 items ($b_{jkF} = b_{jkR}$, $a_{jF} = 0.8, 1.2,$ or 1.6, $a_{jR} = 1.8$ when $a_{jF} = 0.8$, $a_{jR} = 2.5$ when $a_{jF} = 1.2$, $a_{jR} = 3.2$ when $a_{jF} = 1.6$)
Reference and Focal Group Sample Size Ratios: 2 Levels
1. Equal reference/focal group sample size ratio: 1:1: 2000:2000 2. Unequal reference/focal group sample size ratio: 4:1 3200:800
Ability Distribution Differences: Two levels * 2 levels of skewness
Ability Distributions: Equal
1. Moderately skewed: reference and focal group were both (N (0,1)), skewness = - 0.75. 2. Standard normal: reference group and focal group were both (N (0,1)).
Ability Distributions: Unequal
1. Moderately skewed: reference group was (N (0,1)) and focal group was (N (- 0.5,1)); skewness is - 0.75. 2. Standard normal: reference group was (N(0,1)) and focal group was (N (- 0.5,1)); no skewness.
The total number of conditions = 72 (9 studied item levels *2 sample size ratios *2 skewness levels * 2 ability differences). 400 replications were done for each condition, giving a total of 28,800 datasets.

Constants

Test Length and Response Categories in Items

All 26 items on the test had four score levels (0,1,2, and 3); the number of levels is equivalent to that of the TIMSS item found in Appendix One.

Item Parameters for Items 1 to 25

The item parameters for the 25 items which were not studied were held constant. The first 24 item parameters were taken from published parameters which were used by French and Miller (1996) in a Monte Carlo study of polytomous DIF detection with logistic regression. A 25th item was added; the parameters for this item complete the pattern of parameters used by French and Miller. The item parameters for the 25 nonstudied items (items 1-25) are shown in Table 5.

Table 5. Parameters for Items 1 - 25

Item #	a	b_1	b_2	b_3
1	0.5	-2.00	0.00	2.00
2	0.5	-2.00	1.00	2.00
3	0.5	-1.00	0.00	1.00
4	0.5	0.00	1.00	2.00
5	0.5	-2.00	-1.00	0.00
6	0.75	-2.00	0.00	2.00
7	0.75	-2.00	1.00	2.00
8	0.75	-1.00	0.00	1.00
9	0.75	0.00	1.00	2.00
10	0.75	-2.00	-1.00	0.00
11	1.00	-2.00	0.00	2.00
12	1.00	-2.00	1.00	2.00
13	1.00	-1.00	0.00	1.00
14	1.00	0.00	1.00	2.00
15	1.00	-2.00	-1.00	0.00
16	1.25	-2.00	0.00	2.00
17	1.25	-2.00	1.00	2.00
18	1.25	-1.00	0.00	1.00
19	1.25	0.00	1.00	2.00
20	1.25	-2.00	-1.00	0.00
21	1.50	-2.00	0.00	2.00
22	1.50	-2.00	1.00	2.00
23	1.50	-1.00	0.00	1.00
24	1.50	0.00	1.00	2.00
25	1.50	-2.00	-1.00	0.00

The Number of Replications

The number of replications influences the precision of parameter estimates (Harwell, 1996). The decision on the number of replications to run in the present study was based largely on the size of confidence intervals around estimates; time and computer resources were also a factor. Table 6 shows confidence intervals around hypothetical estimates of Type I Error and power according to the number of replications performed (adapted from Fleiss, 1981). It can be seen that these intervals are unacceptably wide at 50 and 100 replications, and still wide at 200 replications. They became acceptable at 400 replications. Therefore, 400 replications were performed.

Table 6. Confidence Limits Around Various Proportions (Hypothetical Type I Error or Power) by the Number of Replications

Number of replications	Hypothetical value	Lower bound	Upper bound
50	0.05	0.0000	0.1116
50	0.20	0.0869	0.3131
50	0.50	0.3586	0.6414
100	0.05	0.0164	0.0936
100	0.20	0.1200	0.2800
100	0.50	0.4000	0.6000
200	0.05	0.0192	0.0808
200	0.20	0.1434	0.2566
200	0.50	0.4293	0.5707
400	0.05	0.0282	0.0718
400	0.20	0.1600	0.2400
400	0.50	0.4500	0.5500

Procedure

Criteria for Judging the Procedures: Type I Error and Power

Type I Error and power were calculated for each of the four DIF assessment procedures under various conditions. Significance levels were set at 0.05. Type I Error was defined as the number of times out of 400 that a null-DIF item was falsely rejected at the 0.05 level. For all four procedures, power was defined as the number of times out of 400 that DIF was correctly identified at the 0.05 level. The four procedures (Mantel, GMH, LDFA-overall, and OLR-overall) were compared on Type I Error when there was no DIF in the data, and on power for detecting uniform DIF and nonuniform DIF. LDFA-uniform and nonuniform and OLR-uniform and nonuniform were also compared on their ability to discriminate between uniform and nonuniform DIF. For LDFA-uniform and OLR-uniform, this was assessed by calculating the number of times out of 400 that they met criteria for nonuniform DIF. For LDFA-nonuniform and OLR-nonuniform, this was assessed by calculating the number of times out of 400 that they met criteria for uniform DIF.

OLR

All three components of the OLR procedure were evaluated:

1) OLR overall. Here, the two-degree of freedom χ^2 test of the difference between the fully saturated and the null model had to exceed 5.99 and the total difference in item scores had to exceed 0.03.

2) OLR-uniform. The one-degree of freedom χ^2 test of the difference between the uniform model and the null model had to exceed 3.84 and the mean difference in item scores had to exceed 0.03.

3) OLR-nonuniform.. If the fully saturated model was significant, and uniform DIF was not found, the item was said to have nonuniform DIF.

OLR-overall was evaluated in comparison to Mantel, GMH, and LDFA-overall in terms of Type I Error and power to detect uniform and nonuniform DIF. OLR-uniform and OLR-nonuniform were evaluated in terms of discriminating types of DIF.

When effect size was excluded from the decision process, statistical significance alone was the basis for identifying DIF.

LDFA

All three components of the LDFA procedure were evaluated:

1) LDFA-overall. This two-degree of freedom χ^2 test of the difference between the full model (nonuniform) and the null model had to exceed 5.99 to be significant and the probability of group membership had to exceed 0.16 in order for DIF to be identified. This cut-off was determined from pretests.

2) LDFA- uniform.. Here, the one-degree of freedom χ^2 test of difference between the uniform model and the null model had to exceed 3.84 and the mean probability of group membership had to exceed 0.16.

3) LDFA-nonuniform. In this test, the one-degree of freedom χ^2 test of the difference between the uniform model and the null model had to exceed 3.84 and the probability of group membership had to exceed 0.16. Because the sequential modelling strategy used in this study for LDFA and OLR involves an overall test, followed by tests of the uniform and nonuniform models to discriminate between types of DIF, LDFA-overall was evaluated in comparison to Mantel, GMH, and OLR-overall on Type I Error rates, and power to detect

uniform and nonuniform DIF. LDFA-uniform and LDFA-nonuniform were evaluated in terms of misidentification of the type of DIF (e.g. Type I error for uniform when only nonuniform DIF was present). Data on the Type I Error for null DIF and power of LDFA-uniform and LDFA-nonuniform for uniform and nonuniform DIF are also presented in order to show that the results would be the same regardless of whether LDFA-overall was used or not.

When effect size was excluded from the decision process, the same steps were followed, except that statistical significance alone was the basis for identifying DIF.

The Mantel

When effect size was included, the Mantel χ^2 had to exceed 3.84 to be significant AND the mean difference in item scores between the reference and focal groups had to exceed 0.03.

When effect size was excluded from the decision process, a χ^2 (1, 4000) greater than 3.84 was the sole basis for identifying DIF.

The GMH

GMH is distributed as χ^2 with k - 1 degrees of freedom, where k is the number of response levels in the item (in this case, four). For the analyses with effect size, this χ^2 (3, 4000) value had to exceed 7.81 to reach significance AND the mean total difference in item score had to exceed 0.03 in order to identify DIF in that item.

When effect size was excluded from the decision process, a χ^2 (1, 4000) greater than 7.81 was needed to identify DIF.

Steps in the Simulation Study

For each condition:

1. Ability values for reference and focal groups were randomly generated using the appropriate level of skewness, group ability difference, and sample size ratio.

2. Item responses for each sampled ability value were generated according to specified item parameters using Muraki's (1992) Generalized Partial Credit Model (GPCM), which gives item characteristic curves as functions of latent ability. The GPCM model is analagous to a two-parameter logistic model for ordered response categories; in it the slope (a) and threshold (b) parameters are estimated, the guessing (c) parameter is not estimated.

3. The Mantel, GMH, LDFA, and Ordinal Logistic Regression procedures, with and without effect size measures, were performed on each dataset.

4. For each test (1 GMH, 1 Mantel, 3 LDFA, and 3 Ordinal Logistic Regression), a count was made of whether or not the appropriate value(s) was (were) exceeded with effect size included, and next, without effect size included. These were saved in a file.

5. All of the above steps were repeated 400 times for each combination of independent variables (these combinations are referred to as 'study conditions').

Computer Programs

The data generation program was a modification of PSYCH.FOR (Spray, 1992), which was designed to generate polytomous item responses using the GPCM It was converted to SAS, and linked to programs which performed the four DIF assessment procedures: GENMH.FOR (Spray, 1992), for Generalized Mantel-Haenszel, PSYCH.FOR for the Mantel and LDFA procedures, and the logistic regression procedure in SAS. The DIF assessment programs were all converted to SAS

6.9 by Aylesworth. SAS has several advantages over FORTRAN, including a better data generation seed and higher efficiency. As the program ran, results from each procedure for each replication were saved in an output file for later analysis.

Validation of the Data Generation and DIF Assessment Programs

The validity of this study rests on the accuracy of the computer programs. Therefore, the data generation program and the programs which were used to run the four DIF assessment procedures were checked carefully and thoroughly.

Checking Data Generation Programs. The generated data were examined graphically and statistically to ensure that each study condition matched specifications. These examinations demonstrated that the data generation program worked as it was designed to work. Four examples of different ability*skewness conditions are shown in Table 7; one iteration from each was randomly chosen for presentation. This table clearly shows that the data produced by the program matched study specifications. For example, Condition 4 was designed to have unequal (- 0.5) and negatively skewed ability distributions. These requirements were met; the mean ability of the focal group was almost exactly 0.5 standard deviations lower than that of the reference group (- 0.51); the mean total test score was nearly 7 points lower for the focal group. The ability distributions for both the reference and focal group were moderately skewed (- 0.85 for both).

Checking DIF Assessment Programs. In another set of checks, the accuracy of the SAS programs written for the Mantel, GMH, and LDFA was assessed by thoroughly checking the mathematics and by comparing the results of each SAS program to results obtained from the original FORTRAN program. Because the procedure was changed, output from the SAS program

was not compared to the original SPSS program. However, the mathematics were carefully checked.

Table 7. Examples of Four Different Ability*Skewness Conditions: One Iteration

	Mean Ability (*z*-score)	Mean Total Score (out of 79)	Skewness in Ability Distribution
Condition 1: No ability difference, no skewness	R: 0.01	38.01	- 0.06
	F: - 0.02	37.53	- 0.07
Condition 2: No ability difference, skewness	R: - 0.04	37.65	- 0.86
	F: 0.03	38.40	- 0.70
Condition 3: Ability difference, no skewness	R: 0.01	38.06	0.29
	F: - 0.47	31.70	0.09
Condition 4: Ability difference and skewness	R: - 0.05	38.00	- 0.85
	F: - 0.51	31.40	- 0.85

Final Analyses

Comparing DIF Assessment Procedures

Tables showing the mean (over 400 replications) power and Type I Error of the DIF assessment procedures for each of the seventy-two study conditions were drawn up. Type I Error rates below 0.05 were the ideal; rates above 0.10 were considered unacceptably high. Power below 0.80 was considered to be unacceptably low. These values are somewhat arbitrary, but are reasonable and typical of values used in research. More importance is attached to Type I Error than to power, which is quite reasonable (Cohen, 1977) given the fact that items should only be removed from tests if there is very strong evidence for bias (Roznowski & Reith, 1999).

To compare procedures, the overall power and Type I Error of each procedure (and for LDFA and OLR, each test) for uniform and nonuniform DIF across all conditions were calculated as was the number of times that Type I Error exceeded 0.05, 0.10 and, for power, the number of times that power was below 0.80. In order to compare LDFA and OLR for discriminating types of DIF, Type I Error for the uniform tests was calculated under conditions of nonuniform DIF, and Type I Error of the nonuniform tests was calculated under conditions of uniform DIF.

Evaluating the Relationship between Independent Variables and Performance of the DIF Assessment Procedures

Logistic regression was used to evaluate the effect of the independent variables on Type I Error and power of the Mantel, GMH, LDFA, and OLR procedures. Logistic regression describes the relationship between a dichotomous dependent variable and a set of explanatory variables (Stokes, Davis, & Koch, 1995). Separate logistic regressions were conducted for each procedure (Mantel, GMH, LDFA-overall, and OLR-overall (all with effect size)) and for each DIF condition. Logistic regression analyses were also performed with LDFA-uniform and nonuniform and OLR-uniform and nonuniform in order to evaluate the effect of the independent variables on discriminating between uniform and nonuniform DIF. The dependent variable for each analysis was a count of the number of times the null was rejected over the total number of observations. The independent variables were:

1. Sample size ratio, or SS. This was coded 1 for 4:1 and 0 for 1:1.
2. Studied item discrimination, D. This variable had three levels, low (0.8), medium (1.2), and high (1.6). The lowest level was coded as the reference level while the other two levels were

dummy coded. For the moderate discrimination level of 1.2 (variable D2),1 = an item discrimination of 1.2 and 0 = an item discrimination value of 0.8 or 1.6. For the high item discrimination of 1.6 (variable D3), 1 = an item discrimination of 1.6 and 0 = an item discrimination value of 0.8 or 1.2.

3. Difference in ability distributions, or DA. This variable was coded as 1 when there was a group ability difference of 0.5 and 0 when there was no group ability difference.

4. Skewness, or SK. This variable was coded as 1 when the data were skewed, and 0 when the data were normally distributed.

All main effects and second-order interactions were tested. In order to avoid overfitting, third and fourth order interactions were not tested. Stepwise selection was used to determine which variables would be included in the final model. In stepwise selection, the intercept is entered first. Next, the variable with the largest chi-square value that meets entry criterion ($p <$ 0.05) is entered. At each step, the p-values of all variables in the model were examined; those with p-values of less than 0.15 were removed from the model. In these analyses, the odds ratios (OR) represented the ratio of the odds of rejecting the null when the independent variable under study was 1 to the odds of rejecting the null when the independent variable was 0. Parsimony was a guiding principle in all analyses. Individual effects were considered to be significant only if their χ^2 values were significant and if their odds ratio either exceeded 1.5 or were less than 0.5 (one exception to this occurred when the variable was needed for good model fit). These thresholds are typical of those used in research to ensure that the independent variables are practically significant as well as statistically significant.

Evaluating the Impact of Effect Size

The purpose behind inclusion of the magnitude of difference in item score between groups (effect size) is to provide the user of the procedure with an indication of the importance of the difference. The inclusion of an effect size measure should reduce Type I Error. In order to study the impact of effect size on Type I Error, a comparison was made between Type I Error when effect size was used in the decision rules and Type I Error when decisions were based on statistical significance alone. These comparisons were made for LDFA-overall, OLR-overall, LDFA-uniform, OLR-uniform, the Mantel, and the GMH. An overall mean of Type I Error with and without the inclusion of effect size was also calculated. To test for mean differences in Type I Error rates and power with and without effect size for each procedure, z-tests for differences in proportions (Fleiss, 1981) were performed.

Effect size may also change the power of procedures. Therefore, the above comparisons were also done for power of the LDFA-overall, OLR-overall, the Mantel, and GMH procedures.

CHAPTER IV
RESULTS

This chapter comprises five sections. In the first four sections, Type I error and power results are presented for each procedure. These results include tables and written summaries of the data, and the logistic regression analyses. In the fifth section, results are presented and compared with and without effect size.

The Mantel Procedure

Detecting DIF in Test Items

Type I Error Rates. With the inclusion of the measure of effect size, Type I error rates for the Mantel procedure were extremely low, ranging from 0.008 to 0.043 (see Table 8); the mean was 0.021. None of the values exceeded 0.05. Results of the logistic regression analysis showed that the interaction between high item discrimination and group ability difference (D3AD) was related to increased Type I Error. A model with D3AD (OR = 1.78, Wald $\chi^2 = 12.3$, $p < .001$) included as the sole predictor explained the Type I Error rates of the Mantel (Model $\chi^2 = 11.2$, df = 1, $p < .001$).

Power for Uniform DIF. The Mantel had excellent power for uniform DIF (see Table 8). Power ranged from 0.873 to 1.00 and the mean was 0.983. All values were well above 0.80; power reached 1.00 in twelve of the twenty-four study conditions. Logistic regression analyses demonstrated that the power of the Mantel for uniform DIF was significantly related to three predictors: sample size (SS), moderate item discrimination (D2), and high item discrimination (D3). The power of the Mantel to detect uniform DIF increased with increasing item

discrimination (D2: OR = 28.7, Wald $\chi^2 = 64.8$, $p < .001$) and was highest when item discrimination was high (D3: OR = 172.7, Wald $\chi^2 = 26.4$, $p < .001$). In contrast, the power of the Mantel for uniform DIF decreased when the sample size ratio was 4:1 (SS: OR = 0.172, Wald $\chi^2 = 65.8$, $p < .001$). The overall model fit was excellent ($\chi^2 = 279.5$, df = 4, $p < .001$).

Table 8. Type I Error and Power of the Mantel Procedure for Detecting Uniform DIF

	Equal Abilities (R-*N* (0,1), F-*N* (0,1)					
	Type I Error (No DIF)			**Power (Uniform DIF present)**		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Low (0.8)	Medium (1.2)	High (1.6)
No Skew						
Ratio: 1:1	0.018	0.015	0.010	0.988	1.000	1.000
4:1	0.020	0.020	0.018	0.938	1.000	1.000
Skew						
Ratio 1:1	0.030	0.010	0.028	0.988	1.000	1.000
4:1	0.020	0.020	0.015	0.945	0.995	0.998
	Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)					
	Type I Error			**Power**		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Low (0.8)	Medium (1.2)	High (1.6)
No Skew						
Ratio: 1:1	0.025	0.015	0.043	0.980	1.000	1.000
4:1	0.013	0.025	0.028	0.873	0.995	1.000
Skew						
Ratio: 1:1	0.018	0.008	0.035	0.980	1.000	1.000
4:1	0.025	0.013	0.023	0.913	0.995	1.000

The detection rates for the Mantel by item discrimination, ability, and sample size ratio are shown in Table 9. The Type I Error for the Mantel was relatively the highest when item discrimination was high and ability distributions were unequal (M = 0.032). This effect was reversed at low levels of item discrimination where Type I Error was somewhat higher when ability distributions were equal. The power of the Mantel to detect uniform DIF was somewhat lower when the sample size ratio was 4:1 than when it was 1:1. On the other hand, power was higher when item discrimination was moderate (M = 0.998) than when it was low (M = 0.953) and increased slightly more when item discrimination was high (M = 0.999).

Table 9: Type I Error and Power for Uniform DIF of the Mantel: Item Discrimination by Group Ability Difference and Sample Size Ratio

	Type I Error (No DIF)			**Power (Uniform DIF present)**		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Med. (1.2)	High (1.6)	Low (0.8)	Med. (1.2)	High (1.6)
1:1 Ratio						
Equal	0.024	0.013	0.019	0.988	1.00	1.00
Unequal	0.022	0.011	0.039	0.980	1.00	1.00
4:1 Ratio						
Equal	0.020	0.020	0.017	0.941	0.998	0.999
Unequal	0.019	0.019	0.025	0.893	0.995	1.000

Note: Each entry is averaged across 2 levels of skewness and is based on 800 observations.

Power for Nonuniform DIF. The power of the Mantel procedure for detecting nonuniform DIF is presented below in Table 10. It can be seen that the Mantel had almost no power to detect nonuniform DIF. Power ranged from 0.038 to 0.24, and was well below the acceptable value of 0.80 in all conditions; the mean was 0.096. Therefore, it is not meaningful to

assess the effects of other factors on the power of the Mantel for nonuniform DIF and logistic regression analyses were not run.

Table 10. Type I Error and Power of the Mantel Procedure for Detecting Nonuniform DIF

	Type I Error (No DIF)			Power (Nonuniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Focal 0.8 Ref : 1.8	Focal: 1.2 Ref: 2.5	Focal: 1.6 Ref: 3.2
Equal Abilities (R-*N* (0,1), F-*N* (0,1)						
No skewness						
Ratio: 1:1	0.018	0.015	0.010	0.048	0.038	0.055
4:1	0.020	0.020	0.018	0.103	0.138	0.110
Skewness						
Ratio 1:1	0.030	0.010	0.028	0.048	0.053	0.053
4:1	0.020	0.020	0.015	0.133	0.093	0.115
Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)						
	Type I Error (No DIF)			Power (Nonuniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Focal 0.8 Ref : 1.8	Focal: 1.2 Ref: 2.5	Focal: 1.6 Ref: 3.2
No Skewness						
Ratio: 1:1	0.025	0.015	0.043	0.168	0.063	0.040
4:1	0.013	0.025	0.028	0.243	0.093	0.073
Skewness						
Ratio: 1:1	0.018	0.0008	0.035	0.155	0.065	0.058
4:1	0.025	0.013	0.023	0.223	0.093	0.050

Distinguishing Types of DIF

The Mantel does not model uniform and nonuniform DIF separately. Therefore, it cannot be used to distinguish between different types of DIF.

Summary of Results for the Mantel Procedure

The Mantel procedure (with effect size) had excellent Type I Error control; the mean Type I Error was only 0.021. Results of the logistic regression analyses showed that Type I Error was slightly but significantly increased when item discrimination was high and a small group ability difference existed, although Type I Error always remained below 0.05. The Mantel also had excellent power for uniform DIF ($\underline{M}$ = 0.983); logistic regression showed that power increased as group ability difference increased. However, power decreased slightly when the sample size ratio was 4:1. In all conditions, power was well above the nominal value of 0.80. As expected, the Mantel procedure had no power to detect nonuniform DIF. It also cannot be used to discriminate between uniform and nonuniform DIF.

The GMH Procedure

Detecting DIF in Test Items

Type I Error. The Type I Error for the GMH procedure (with effect size) ranged from 0.03 to 0.065 (see Table 11; repeated in 13); the mean Type I Error was 0.0436. The Type I Error for the GMH procedure exceeded the nominal rate of 0.05 in only four out of twenty-four conditions; all values were below 0.10. The logistic regression analysis indicated that the Type I Error of the GMH was significantly related to only one variable: the interaction between group ability difference and high item discrimination (D3AD). The Type I Error for the GMH was

somewhat higher when item discrimination was high and when there were group ability differences (D3AD: OR = 1.36, Wald $\chi^2 = 6.15, p < .01$).

Power for Uniform DIF. The GMH had excellent power for uniform DIF, which ranged from 0.937 to 1.00 (see Table 11). All power values were well above 0.80 in all 24 uniform DIF conditions, and the overall mean power was 0.983.

Table 11. Type I Error and Power of the GMH Procedure for Detecting Uniform DIF

	Type I Error (No DIF)			Power (Uniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Low (0.8)	Medium (1.2)	High (1.6)
Equal Abilities (R-*N* (0,1), F-*N* (0,1)						
No skewness						
Ratio: 1:1	0.033	0.040	0.030	0.99	1.000	1.000
4:1	0.033	0.045	0.048	0.938	1.000	1.000
Skewness						
Ratio 1:1	0.053	0.035	0.053	0.988	1.000	1.000
4:1	0.048	0.055	0.048	0.945	0.998	1.000
Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)						
	Type I Error			Power		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Low (0.8)	Medium (1.2)	High (1.6)
No skewness						
Ratio: 1:1	0.038	0.038	0.065	0.980	1.000	1.000
4:1	0.038	0.045	0.048	0.875	0.995	0.993
Skewness						
Ratio: 1:1	0.040	0.023	0.058	0.980	1.000	1.000
4:1	0.048	0.038	0.050	0.913	0.995	1.000

Three variables were significantly related to the power of the GMH for detecting uniform DIF: moderate item discrimination, (D2), high item discrimination (D3), and sample size ratio (SS) (Model $\chi^2 = 411.3$, df = 4, $p < .001$). As shown in Table 12, the probability of correctly rejecting the null hypothesis was somewhat higher when item discrimination was moderate than when it was low (D2: OR = 34, $\chi^2 = 60$, $p < .001$) and slightly higher yet when item discrimination was high (D3: OR = 170.5, $\chi^2 = 26.2$, $p < .001$). The mean power of the GMH for uniform DIF was 0.95 when item discrimination was low, 0.999 when item discrimination was moderate, and 1.00 when item discrimination was high. The 4:1 sample size ratio resulted in slightly decreased power to detect uniform DIF; mean power was 0.99 with equal sample sizes and 0.97 with unequal sample sizes (see Table 12).

Table 12. Type I Error and Power of GMH for Uniform DIF: Sample Size Ratio * Item Disc.

	Type I Error			**Power (Uniform DIF Present)**		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Low (0.8)	Medium (1.2)	High (1.6)
1:1 ratio	0.041	0.034	0.052	0.98	1.000	1.000
4:1 ratio	0.042	0.046	0.049	0.917	0.997	1.000

Note: values are averaged across two levels of ability distribution and two levels of skewness and represent 1600 observations.

Power for Nonuniform DIF. The GMH procedure (with effect size) had very high power for detecting nonuniform DIF. Power ranged from 0.748 to 1.00 (see Table 13); the overall mean power was 0.934. Only one value was below 0.80. The logistic regression analysis demonstrated that the power of the GMH for nonuniform DIF was significantly related to six variables: moderate item discrimination (D2), high item discrimination (D3), skewness (SK), sample size ratio (SS),

moderate item discrimination *x* sample size ratio (D2SS), and ability distribution skewness (ADSK).

Table 13. Type I Error and Power of the GMH for Detecting Nonuniform DIF

	Type I Error (No DIF)			Power (Nonuniform DIF present)		
Equal Abilities (R-*N* (0,1), F-*N* (0,1)						
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Focal 0.8 Ref : 1.8	Focal: 1.2 Ref: 2.5	Focal: 1.6 Ref: 3.2
No Skew						
Ratio: 1:1	0.033	0.040	0.030	0.985	0.863	1.000
4:1	0.033	0.045	0.048	0.953	0.900	1.000
Skew						
Ratio 1:1	0.053	0.035	0.053	0.975	0.748	0.990
4:1	0.048	0.055	0.048	0.955	0.810	0.990
Unequal Abilities (R-*N* (0,1), F-*N* (-0.5,1)						
	Type I Error (No DIF)			Power (Nonuniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Focal 0.8 Ref : 1.8	Focal: 1.2 Ref: 2.5	Focal: 1.6 Ref: 3.2
No Skew						
Ratio: 1:1	0.038	0.038	0.065	0.988	0.875	1.000
4:1	0.038	0.045	0.048	0.953	0.850	0.993
Skew						
Ratio: 1:1	0.040	0.023	0.058	0.973	0.850	0.998
4:1	0.048	0.038	0.050	0.933	0.838	0.990

The model fit the data well (Model $\chi^2 = 731.1$, df = 6, $p < .001$). As shown in Table 13, the relationship between power of the GMH for nonuniform DIF and item discrimination was somewhat U-shaped; power was lowest when item discrimination was moderate (D2: OR = 0.103, Wald $\chi^2 = 155.4$, $p < .001$), but was highest when item discrimination was high (D3: OR = 7.48, Wald $\chi^2 = 56.24$, $p < .001$). Power was also lower when the sample size ratio was 4:1 than when it was 1:1 (OR = 0.47, Wald $\chi^2 = 12.8$, $p < .001$). The interaction between moderate discrimination and sample size ratio meant that power for nonuniform DIF was higher when item discrimination was moderate and sample size ratio was 4:1 (D2SS: OR = 3.0, Wald $\chi^2 = 24.5$, $p < .001$). Skewness in the ability distributions resulted in somewhat lower power (SK: OR = 0.49, Wald $\chi^2 = 43$, $p < .001$). The interaction between skewness and ability distribution differences meant that the probability of correctly rejecting the null was somewhat higher when the data were skewed and when there was an ability difference (OR = 1.69, Wald $\chi^2 = 14.8$, $p < .001$).

Classifying DIF

Although the GMH procedure has excellent power for detecting both uniform and nonuniform DIF, it only has one overall test, and cannot discriminate between types of DIF.

Summary of Results for the GMH Procedure

The GMH showed good Type I Error control; the mean Type I Error was 0.044; only five values were above 0.05, and none were above 0.10. The logistic regression analyses indicated that the Type I Error of the GMH was increased slightly when item discrimination was high and there was a difference in group ability distribution. The GMH had excellent power for uniform DIF ($\underline{M}$ = 0.983). Power for uniform DIF was increased with high item discrimination, but decreased when the sample size ratio was 4:1. The GMH also had very high power for nonuniform DIF ($\underline{M}$ =

0.934). Results of the logistic regression analysis showed that power for nonuniform DIF was related to item discrimination, sample size ratio, skewness, moderate item discrimination * sample size ratio, and group ability difference * skewness. Power of the GMH for nonuniform DIF had a U-shaped relationship to item discrimination and decreased with a high sample size ratio and with skewed data.

The LDFA Procedure

Detecting DIF in Test Items

Type I Error for LDFA-Overall. LDFA-overall displayed moderately good control over Type I Error; rates ranged from 0.020 to 0.083 (see Table 14); the mean was 0.0496. The Type I Error rates for LDFA-overall exceeded the nominal value of 0.05 in thirteen out of 24 conditions; no values exceeded 0.10. Results of the logistic regression analysis indicated that the interaction between high item discrimination and group ability difference (D3AD) was related to Type I Error of the LDFA procedure. When item discrimination was high (1.6), and there was a difference in group abilities, the probability of Type I Error was increased by 1.5 times over all of the other conditions (D3AD: OR = 1.52, $p < .01$). The overall model fit was good (Model $\chi^2 = 28.9$, $p < .01$).

Uniform DIF. LDFA-overall had excellent power for uniform DIF; this ranged from 0.915 to 1.00 (see Table 14). Power was well above the nominal value of 0.80 in all cases; the overall mean was 0.990. This high power is excellent, but did mean that logistic regression analyses could not be run due to quasicomplete separation of the data points. Quasicomplete separation means that the data are partially overlapping rather than completely separated as they should be; in this case the null was falsely accepted only 103 times and was rejected 9497 times. When

quasicomplete separation occurs, the maximum likelihood estimates are not valid (SAS Institute, 1995). Therefore, logistic regression results are not reported for the LDFA-overall for the detection of uniform DIF.

Table 14. Type I Error and Power of LDFA-Overall for Detecting Uniform DIF

	Type I Error (No DIF)			Power (Uniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Low (0.8)	Medium (1.2)	High (1.6)
Equal Abilities (R-*N* (0,1), F-*N* (0,1)						
No Skew						
Ratio: 1:1	0.050	0.048	0.053	0.993	1.000	1.000
4:1	0.070	0.048	0.048	0.960	1.000	1.000
Skew						
Ratio 1:1	0.050	0.063	0.045	0.990	1.000	1.000
4:1	0.053	0.058	0.043	0.955	1.000	1.000
Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)						
	Type I Error			Power		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Low (0.8)	Medium (1.2)	High (1.6)
No Skew						
Ratio: 1:1	0.050	0.035	0.083	0.988	1.000	1.000
4:1	0.035	0.023	0.043	0.915	1.000	1.000
Skew						
Ratio: 1:1	0.030	0.035	0.088	0.995	1.000	1.000
4:1	0.060	0.030	0.050	0.950	0.998	1.000

Power for Nonuniform DIF. The power of LDFA-overall for nonuniform DIF was strongly related to item discrimination; it was adequate only when item discrimination was low (see Table 15). When item discrimination was low, the power of the LDFA-overall for nonuniform DIF ranged from 0.798 to 0.963 (see Table 15). LDFA-overall had moderately low power for nonuniform DIF when item discrimination was moderate; under these conditions power ranged from 0.418 to 0.670. Power was extremely low when item discrimination was high, ranging from 0.183 to 0.300. Results of the logistic regression indicated that five independent variables were significant predictors of the power of LDFA-overall to detect nonuniform DIF: moderate item discrimination (D2), high item discrimination (D3), sample size ratio (SS), an interaction between moderate item discrimination and sample size ratio (D2SS) and the interaction between high item discrimination and sample size ratio (Model $\chi^2 = 2651$, df = 7, $p < .001$). The main effect of item discrimination has already been described: as item discrimination increased, the power of the LDFA-overall for nonuniform DIF decreased markedly (D2: OR = 0.09, Wald χ^2 = 474.5,

$p < .001$: D3: OR = 0.021 Wald $\chi^2 = 949.7$, $p < .001$). Sample size ratio also had a main effect on power; power was generally lower with 4:1 ratios (SS: OR = 0.36,Wald $\chi^2 = 77.8$, $p < .001$). However, there was an interaction between item discrimination and sample size ratio. At low and moderate levels of item discrimination, power was higher with equal sample size; this was reversed at high levels of discrimination (D2SS: OR = 2.2, Wald $\chi^2 = 33.1$, $p < .001$; D3SS: OR = 3.1 Wald $\chi^2 = 61.95$, $p < .001$). However, power was inadequate at high levels of item discrimination.

Table 15. Type I Error and Power of LDFA-Overall for Detecting Nonuniform DIF

	Type I Error (No DIF)			Power (Nonuniform DIF present)		
Equal Abilities (R-*N* (0,1), F-*N* (0,1)						
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Focal 0.8 Ref : 1.8	Focal: 1.2 Ref: 2.5	Focal: 1.6 Ref: 3.2
No Skew						
Ratio: 1:1	0.050	0.048	0.053	0.963	0.670	0.290
4:1	0.070	0.048	0.048	0.825	0.565	0.300
Skew						
Ratio 1:1	0.050	0.063	0.045	0.918	0.468	0.183
4:1	0.053	0.058	0.043	0.795	0.418	0.223
Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)						
	Type I Error (No DIF)			Power (Nonuniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Focal 0.8 Ref : 1.8	Focal: 1.2 Ref: 2.5	Focal: 1.6 Ref: 3.2
No Skew						
Ratio: 1:1	0.050	0.035	0.083	0.945	0.573	0.238
4:1	0.035	0.023	0.043	0.875	0.530	0.280
Skew						
Ratio: 1:1	0.030	0.035	0.088	0.890	0.488	0.260
4:1	0.060	0.030	0.050	0.798	0.445	0.235

The mean Type I Error and power for nonuniform DIF of LDFA-overall by item discrimination and sample size ratio are shown in Table 16. Here, the strong effect of item discrimination on power for nonuniform DIF is clearly shown; power decreased from a mean of 0.86 with low discrimination to 0.52 under moderate discrimination and to 0.25 with high

discrimination. Furthermore, power was generally somewhat lower when the sample sizes were unequal; this was reversed at the highest level of item discrimination.

Table 16. Type I Error and Power of LDFA-overall for Detecting Nonuniform DIF; Item Discrimination and Sample Size Ratio

	Type I Error (No DIF)			Power (Nonuniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Focal 0.8 Ref : 1.8	Focal: 1.2 Ref: 2.5	Focal: 1.6 Ref: 3.2
1:1 Ratio	0.045	0.045	0.067	0.929	0.550	0.242
4:1 Ratio	0.055	0.039	0.046	0.823	0.490	0.260

Note: Rates are averaged across two levels of group ability difference and two levels of skewness and represent 1600 observations.

Type I Error for LDFA-uniform. The Type I Error rate for LDFA-uniform when there was no DIF ranged from 0.005 to 0.050; the mean was 0.0232 (see Table 17). In all cases, the sum of Type I Error rates for LDFA-uniform and LDFA-nonuniform equaled those of LDFA-overall.

Uniform DIF. Power values for LDFA-uniform ranged from 0.908 to 1.00; these values are well above the minimum acceptable value for all combinations of skewness, sample size ratio, item discrimination, and group ability difference (see Table 17). The mean power of LDFA-uniform for uniform DIF was 0.987. This mean is just slightly below that of LDFA-overall; the power of LDFA-uniform plus the Type I Error of LDFA-nonuniform (when uniform is present) equals the power of LDFA-overall for nonuniform DIF.

Because the main focus was on LDFA-overall, logistic regression analyses were not performed for LDFA-uniform except for differentiating between uniform and nonuniform DIF.

Table 17. Type I Error and Power of LDFA-uniform for Uniform DIF

	Type I Error (No DIF)			Power (Uniform DIF present)		
Equal Abilities (R-*N* (0,1), F-*N* (0,1)	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Low (0.8)	Medium (1.2)	High (1.6)
No Skew						
Ratio: 1:1	0.030	0.020	0.020	0.993	1.000	1.000
4:1	0.020	0.023	0.020	0.945	0.998	1.000
Skew						
Ratio 1:1	0.030	0.028	0.030	0.990	1.000	1.000
4:1	0.020	0.013	0.010	0.948	0.998	1.000
Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)	Type I Error (No DIF)			Power (Uniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Low (0.8)	Medium (1.2)	High (1.6)
No Skew						
Ratio: 1:1	0.025	0.018	0.050	0.988	1.000	1.000
4:1	0.025	0.005	0.020	0.908	1.000	1.000
Skew						
Ratio: 1:1	0.025	0.018	0.055	0.995	1.000	1.000
4:1	0.018	0.013	0.023	0.938	0.998	1.000

Type I Error: LDFA-nonuniform. Type I Error rates for LDFA-nonuniform when there was no DIF ranged from 0.005 to 0.333 (see Table 18); the mean was 0.0264.

Power for Nonuniform DIF. Power of LDFA-nonuniform for nonuniform DIF followed a similar pattern to that of LDFA-overall. Power for nonuniform DIF ranged from 0.13 to 0.91, and

only reached the desired level of 0.80 under two of twenty-four conditions (see Table 18). As with LDFA-overall, power of LDFA-nonuniform was strongly affected by item discrimination. Power ranged from 0.66 to 0.908 when item discrimination was low (M = 0.77).

Table 18. Power and Type I Error of LDFA-nonuniform for Nonuniform DIF

	Equal Abilities (R-*N* (0,1), F-*N* (0,1)					
	Type I Error (No DIF)			**Power (Nonuniform DIF present)**		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Focal 0.8 Ref : 1.8	Focal: 1.2 Ref: 2.5	Focal: 1.6 Ref: 3.2
No Skew						
Ratio: 1:1	0.020	0.028	0.033	0.908	0.630	0.248
4:1	0.050	0.025	0.028	0.783	0.493	0.253
Skew						
Ratio 1:1	0.020	0.035	0.015	0.868	0.425	0.130
4:1	0.033	0.045	0.033	0.730	0.375	0.168
	Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)					
	Type I Error (No DIF)			**Power (Nonuniform DIF present)**		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Focal 0.8 Ref : 1.8	Focal: 1.2 Ref: 2.5	Focal: 1.6 Ref: 3.2
No Skew						
Ratio: 1:1	0.025	0.018	0.033	0.773	0.515	0.208
4:1	0.010	0.018	0.023	0.705	0.463	0.223
Skew						
Ratio: 1:1	0.005	0.018	0.033	0.738	0.415	0.203
4:1	0.043	0.020	0.023	0.660	0.375	0.168

When item discrimination was moderate, power ranged from 0.375 to 0.63 (M = 0.46). Finally, when item discrimination was high, it ranged from 0.13 to 0.25 (M = 0.200).

Discriminating Types of DIF

Type I Error rates for LDFA-uniform when only nonuniform DIF was present ranged from excellent to moderately poor, ranging from 0.03 to 0.173 (see Table 19); the mean was 0.072. Sixteen out of twenty-four values exceeded the nominal value of 0.05. The results of the logistic regression analysis indicated that three variables were related to the probability of misidentifying nonuniform DIF as uniform DIF: these were group ability difference (AD), ability difference *x* moderate item discrimination, and ability difference *x* high item discrimination (Model $\chi^2 = 182$, $p < .001$). As seen in Table 19, nonuniform DIF was more likely to be misidentified as uniform DIF when there was a difference in group ability (AD: OR = 3.52, Wald $\chi^2 = 61.95$, $p < .001$). The interaction between item discrimination and ability difference meant that the probability of misidentification was highest when item discrimination was low, and there was a difference in group ability distribution (M = 0.159). This probability decreased with moderate item discrimination and unequal abilities (M = 0.067: OR = 0.382, Wald $\chi^2 = 63.1$, $p < .001$) and decreased further with high item discrimination and unequal abilities (M = 0.0535: OR = 0.299, Wald $\chi^2 = 85.2$, $p < .001$).

Table 19. Type I Error for LDFA-uniform when Nonuniform DIF was Present

	Equal Abilities (R-*N* (0,1), F-*N* (0,1)		
	Studied Item Discrimination		
	Focal 0.8, Ref : 1.8	Focal: 1.2, Ref: 2.5	Focal: 1.6, Ref: 3.2
No Skew			
Ratio: 1:1	0.055	0.040	0.043
4:1	0.043	0.073	0.048
Skew			
Ratio 1:1	0.050	0.043	0.053
4:1	0.065	0.043	0.055
	Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)		
	Studied Item Discrimination		
	Focal 0.8, Ref : 1.8	Focal: 1.2, Ref: 2.5	Focal: 1.6, Ref: 3.2
No Skew			
Ratio: 1:1	0.173	0.058	0.030
4:1	0.170	0.068	0.058
Skewness			
Ratio: 1:1	0.153	0.073	0.058
4:1	0.138	0.070	0.068

Type I Error for LDFA-nonuniform when Only Uniform DIF was Present. The Type I Error rates for LDFA-nonuniform when only uniform DIF was present were quite low, ranging from 0 to 0.015 (see Table 20); the mean was 0.0021. No values exceeded the 0.05 threshold. Because misidentification of uniform DIF as nonuniform DIF was so rare, there was quasicomplete separation among the data points and results of logistic regression are not reported.

Table 20. Type I Error for LDFA-nonuniform when Uniform DIF was Present

Equal Abilities (R-*N* (0,1), F-*N* (0,1)			
	Studied Item Discrimination		
	Low (0.8)	Moderate (1.2)	High (1.6)
No Skew			
Ratio: 1:1	0.000	0.000	0.000
4:1	0.015	0.003	0.000
Skew			
Ratio 1:1	0.000	0.000	0.000
4:1	0.008	0.003	0.000
Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)			
	Studied Item Discrimination		
	Low (0.8)	Moderate (1.2)	High (1.6)
No Skew			
Ratio: 1:1	0.000	0.000	0.000
4:1	0.008	0.000	0.000
Skew			
Ratio: 1:1	0.000	0.000	0.000
4:1	0.013	0.000	0.000

Summary of Results for LDFA

LDFA-overall had moderately good Type I Error control; thirteen values exceeded 0.05, but none exceeded 0.10. The Type I Error rate was increased when there was high item discrimination and group ability differences. LDFA-overall had excellent power for uniform DIF; the mean was 0.990. On the other hand, results for LDFA-overall and LDFA-nonuniform for nonuniform DIF were disappointing; mean power was only 0.55 for LDFA-overall and only 0.48 for LDFA-nonuniform. Results of the logistic regression analysis indicated that power of the

LDFA for nonuniform DIF was strongly related to item discrimination; it decreased sharply as item discrimination increased. Power also decreased with a 4:1 sample size ratio. There was an interaction between sample size and item discrimination. At low and moderate levels of item discrimination, power was lowest when sample sizes were unequal; this was reversed when item discrimination was high.

In terms of discriminating between types of DIF, Type I Error rates for LDFA-uniform ranged from excellent to moderately poor when there was uniform DIF; Type I Error was highest when item discrimination was low. LDFA-nonuniform had very low Type I Error.

The OLR Procedure

Detecting DIF in Test Items

Type I Error Rates. Type I Error rates for OLR-overall were quite low, ranging from 0.025 to 0.068 (see Table 21); the mean was 0.0435. Type I Error rates for OLR-overall exceeded 0.05 in five of the 24 combinations of sample size ratio, item discrimination, skewness, and ability distribution; none exceeded 0.10. It is difficult to find any clear patterns in Type I Error in Table 21. This observation is supported by the fact that in the logistic regression analysis, none of the independent variables were significantly related to the probability of Type I Error for the OLR procedure.

Power for Uniform DIF. OLR-overall had high power for detecting uniform DIF; power ranged from 0.803 to 1.00 (see Table 21); the mean was 0.963. The power of OLR-overall for uniform DIF under all conditions exceeded the minimum acceptable level of 0.80 (see Table 21). The power of OLR-overall for uniform DIF was significantly related to three variables: moderate

item discrimination, high item discrimination, and sample size ratio. The overall fit of the model to the data was very good (Model $\chi^2 = 711.3$, df = 4, $p < .001$).

Table 21. Type I Error and Power of OLR-overall for Detecting Uniform DIF

	Type I Error (No DIF)			Power (Uniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low(0.8)	Med (1.2)	High (1.6)	Low (0.8)	Med. (1.2)	High (1.6)
Equal Abilities (R-*N* (0,1), F-*N* (0,1)						
No Skew						
Ratio: 1:1	0.038	0.035	0.025	0.968	1.000	1.000
4:1	0.050	0.048	0.045	0.850	0.978	1.000
Skew						
Ratio 1:1	0.043	0.040	0.025	0.968	1.000	1.000
4:1	0.050	0.045	0.043	0.880	0.988	0.995
Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)						
	Type I Error (No DIF)			Power (Uniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Med (1.2)	High (1.6)	Low (0.8)	Med (1.2)	High (1.6)
No Skew						
Ratio: 1:1	0.068	0.043	0.055	0.948	1.000	1.000
4:1	0.025	0.040	0.045	0.803	0.983	1.000
Skew						
Ratio: 1:1	0.058	0.053	0.040	0.945	1.000	1.000
4:1	0.055	0.033	0.043	0.825	0.978	1.000

As shown in Table 22, the power of OLR-overall for uniform DIF increased with studied item discrimination; mean power was 0.896 when item discrimination was low (0.8), 0.992 when

item discrimination was moderate (D2: OR = 12.6, Wald $\chi^2 = 171.1, p < .001$) and 0.999 when item discrimination was high (D3: OR = 190.9, Wald $\chi^2 = 54.7, p < .001$). On the other hand, power for uniform DIF was lower when the ratio of subjects in the reference group to that of subjects in the focal group was 4:1; the mean was 0.985 when sample sizes in the reference and focal groups were equal, and decreased to 0.941 when sample sizes were unequal (SS: OR = 0.21, Wald $\chi^2 = 127.6, p < .001$).

Table 22. Power of OLR-overall for Uniform DIF: Item Disc. and Sample Size Ratio

	Power (Uniform DIF present)		
	Studied Item Discrimination		
	Low (0.8)	Med. (1.2)	High (1.6)
1:1	0.957	1.000	1.000
4:1	0.840	0.982	0.998

Note: Mean results across group ability distribution and skewness; each represents 1600 replications.

Power for Nonuniform DIF. OLR-overall had very high power for detecting nonuniform DIF; values ranged from 0.75 to 1.00. Only one value was below 0.80 (see Table 23) and the mean was 0.934. The results of the logistic regression indicate that seven independent variables were significantly related to power for nonuniform DIF: moderate and high item discrimination, sample size ratio, skewness, the interaction between moderate item discrimination and sample size ratio, the interaction between high item discrimination and skewness, and the interaction between ability distribution and skewness. As seen in Table 23, the power of the OLR for nonuniform DIF had a slight U-shaped relationship to item discrimination; it was high at low levels of item discrimination (M = 0.964), but decreased significantly when item discrimination was moderate (M

= 0.854: D2: OR = 0.1, D2: OR = 12.6, Wald $\chi^2 = 171.1, p < .001$). Power was highest at high levels of item discrimination (M = 0.997: D3: OR = 47.7, D2: OR = 12.6, Wald $\chi^2 = 14.8, p < .001$).

Table 23. Type I Error and Power of the OLR-Overall for Detecting Nonuniform DIF

	Type I Error (No DIF)			Power (Nonuniform DIF present)		
Equal Abilities (R-*N* (0,1), F-*N* (0,1)						
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Focal 0.8 Ref : 1.8	Focal: 1.2 Ref: 2.5	Focal: 1.6 Ref: 3.2
No Skew						
Ratio: 1:1	0.038	0.035	0.025	0.985	0.863	1.000
4:1	0.050	0.048	0.045	0.953	0.900	1.000
Skew						
Ratio 1:1	0.043	0.040	0.025	0.975	0.748	0.990
4:1	0.050	0.045	0.043	0.955	0.810	0.990
Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)						
	Type I Error (No DIF)			Power (Nonuniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Focal 0.8 Ref : 1.8	Focal: 1.2 Ref: 2.5	Focal: 1.6 Ref: 3.2
No Skew						
Ratio: 1:1	0.068	0.043	0.055	0.988	0.875	1.000
4:1	0.025	0.040	0.045	0.953	0.850	0.998
Skew						
Ratio: 1:1	0.058	0.053	0.040	0.973	0.850	0.998
4:1	0.055	0.033	0.043	0.933	0.838	0.990

Power was lower when the sample size ratio was 4:1 than when it was 1:1 (SS: OR = 0.48,Wald $\chi^2 = 12.3, p < .001$). However, there was an interaction between item discrimination and sample

size ratio; at the moderate level of item discrimination, power was slightly higher when sample size was 4:1 (D2SS: OR = 2.9,Wald $\chi^2 = 23.5, p < .001$). The power of the OLR for nonuniform DIF was also significantly related to skewness; power was lower when the data were skewed (SK: OR = 0.5, Wald $\chi^2 = 38.9, p < .001$). The interaction between skewness and item discrimination meant that with high item discrimination, power was slightly lower when the data were skewed (D3SK: OR = 0.12, Wald $\chi^2 = 4.3, p < .05$). Finally, the interaction between ability distribution and skewness meant that power was slightly higher when group ability differences were present, and when the data were skewed (ADSK: OR = 1.7, Wald $\chi^2 = 14.4, p < .001$).

Type I Error for OLR-uniform: Null DIF. Type I Error rates for OLR-uniform when there was no DIF in the data were quite low, ranging from 0.005 to 0.04 (see Table 24); the mean Type I Error rate for OLR-uniform was 0.024.

Power of OLR-uniform for Uniform DIF. As shown in Table 24, OLR-uniform also had excellent power for uniform DIF; values ranged from 0.795 to 1.00 ($\underline{M}$ = 0.96).

Table 24. Type I Error and Power of the OLR-uniform for Detecting Uniform DIF

	Type I Error (No DIF)			Power (Uniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low(0.8)	Med (1.2)	High (1.6)	Low (0.8)	Med. (1.2)	High (1.6)
Equal Abilities (R-*N* (0,1), F-*N* (0,1)						
No Skew						
Ratio: 1:1	0.020	0.018	0.005	0.963	1.000	1.000
4:1	0.013	0.025	0.020	0.843	0.978	1.000
Skew						
Ratio 1:1	0.025	0.015	0.015	0.968	1.000	1.000
4:1	0.018	0.025	0.010	0.875	0.988	0.995
Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)						
	Type I Error (No DIF)			Power (Uniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low(0.8)	Med (1.2)	High (1.6)	Low (0.8)	Med. (1.2)	High (1.6)
No Skew						
Ratio: 1:1	0.028	0.020	0.045	0.945	1.000	1.000
4:1	0.010	0.018	0.023	0.795	0.983	1.000
Skew						
Ratio: 1:1	0.033	0.013	0.018	0.945	1.000	1.000
4:1	0.025	0.015	0.028	0.825	0.978	1.000

Type I Error for OLR-Nonuniform: Null DIF. As shown in Table 25, Type I Error rates for OLR-nonuniform when there was no DIF in the data were quite low, ranging from 0.01 to 0.04 ($\underline{M}$ = 0.023). As with LDFA, the Type I Error rates for OLR-uniform and OLR-nonuniform summed to those of OLR-overall.

Table 25. Type I Error and Power of OLR-nonuniform for Detecting Nonuniform DIF

	Type I Error (No DIF)			Power (Nonuniform DIF present)		
Equal Abilities (R-*N* (0,1), F-*N* (0,1)						
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Focal 0.8 Ref : 1.8	Focal: 1.2 Ref: 2.5	Focal: 1.6 Ref: 3.2
No Skew						
Ratio: 1:1	0.018	0.018	0.020	0.978	0.850	0.983
4:1	0.038	0.023	0.025	0.848	0.835	0.940
Skew						
Ratio 1:1	0.018	0.025	0.010	0.973	0.748	0.988
4:1	0.033	0.020	0.033	0.855	0.780	0.933
Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5, 1)						
	Type I Error (No DIF)			Power (Nonuniform DIF present)		
	Studied Item Discrimination			Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)	Focal 0.8 Ref : 1.8	Focal: 1.2 Ref: 2.5	Focal: 1.6 Ref: 3.2
No Skew						
Ratio: 1:1	0.040	0.023	0.010	0.933	0.873	0.998
4:1	0.015	0.023	0.023	0.650	0.725	0.903
Skew						
Ratio: 1:1	0.025	0.040	0.023	0.920	0.840	0.990
4:1	0.030	0018	0.015	0.705	0.723	0.913

Power of OLR-nonuniform for Nonuniform DIF. OLR-nonuniform also had high power for detecting nonuniform DIF. Power ranged from 0.64 to 0.997; only six values were below 0.80 (see Table 25). The mean power across all study conditions was 0.89.

Discriminating between Uniform and Nonuniform DIF

Type I Error rates for OLR-uniform when the data contained only nonuniform DIF were very low to moderate, ranging from 0 to 0.30 (see Table 26); the mean was 0.064. Twelve out of twenty-four values exceeded 0.05. This form of Type I Error was affected strongly by group sample size ratio; Type I Error was well under control when sample sizes were equal, but inflated when the sample size ratio was 4:1 (see Table 26).

Table 26. Type I Error for OLR-uniform when Nonuniform DIF was Present

Equal Abilities (R-*N* (0,1), F-*N* (0,1)			
	Studied Item Discrimination		
	Focal 0.8, Ref : 1.8	Focal: 1.2, Ref: 2.5	Focal: 1.6, Ref: 3.2
No Skew			
Ratio: 1:1	0.008	0.013	0.018
4:1	0.105	0.065	0.060
Skew			
Ratio 1:1	0.003	0.000	0.003
4:1	0.100	0.030	0.058
Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)			
	Studied Item Discrimination		
	Focal 0.8, Ref : 1.8	Focal: 1.2, Ref: 2.5	Focal: 1.6, Ref: 3.2
No Skew			
Ratio: 1:1	0.055	0.003	0.003
4:1	0.303	0.125	0.095
Skew			
Ratio: 1:1	0.053	0.010	0.008
4:1	0.228	0.115	0.078

Results of the logistic regression analysis indicated that the Type I Error of OLR-uniform for discriminating between uniform and nonuniform DIF varied systematically by several factors: sample size ratio, item discrimination, group ability difference, and the interaction between item discrimination and group ability difference (Model $\chi^2 = 727.5$, df =6, $p < .001$). The probability of misidentifying nonuniform DIF as uniform DIF was significantly higher (see Table 27) when the sample size ratio was 4:1 than when it was 1:1 (OR for SS = 9.3, Wald $\chi^2 = 291.6$, $p < .001$). Item discrimination was also related; the probability of misidentifying nonuniform DIF as uniform DIF was significantly lower when item discrimination was moderate (D2: OR = 0.37, $\chi^2 = 88.9$, $p < .001$), and was somewhat lower when item discrimination was high (D3: OR = 0.56,Wald $\chi^2 = 11.1$, $p < .001$). The Type I Error of OLR-uniform for distinguishing between uniform and nonuniform DIF was also related to ability differences; when ability differences were present, the probability of misidentifying nonuniform DIF as uniform DIF was three times higher (AD: OR = 3.2, Wald $\chi^2 = 114.61$, $p < .001$). There was an interaction between group ability difference and item discrimination; the probability of misidentifying nonuniform DIF as uniform DIF decreased when there were group ability differences and high item discrimination (D3AD: OR = 0.42, Wald $\chi^2 = 16.4$, $p < .001$).

Table 27. Type I Error for OLR-uniform when Nonuniform DIF was Present: Sample Size Ratio, Item Discrimination, and Ability Distribution

Equal Abilities (R-*N* (0,1), F-*N* (0,1)			
		Studied Item Discrimination	
	Focal 0.8, Ref : 1.8	Focal: 1.2, Ref: 2.5	Focal: 1.6, Ref: 3.2
Ratio: 1:1	0.0055	0.0065	0.0105
4:1	0.1025	0.0475	0.059
Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)			
		Studied Item Discrimination	
	Focal 0.8, Ref : 1.8	Focal: 1.2, Ref: 2.5	Focal: 1.6, Ref: 3.2
Ratio: 1:1	0.054	0.0065	0.0055
4:1	0.2655	0.12	0.0865

Note: Each entry is averaged across two levels of skewness and represents 800 replications

The Type I Error rates for OLR-nonuniform when the data contained only uniform DIF were extremely low, ranging from 0 to 0.01 (see Table 28). In fact, nineteen of twenty-four Type I Error values for OLR-nonuniform were equal to 0 ($\underline{M} = 0.004$). The logistic regression analysis could not be performed due to this extremely low Type I Error rate and consequent quasicomplete separation of the data points (only 11 events out of 9600 trials).

Table 28. Type I Error for OLR-nonuniform when Uniform DIF was Present

Equal Abilities (R-*N* (0,1), F-*N* (0,1)			
	Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)
No Skew			
Ratio: 1:1	0.005	0.000	0.000
4:1	0.008	0.000	0.000
Skew			
Ratio 1:1	0.000	0.000	0.000
4:1	0.005	0.000	0.000
Equal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)			
	Studied Item Discrimination		
	Low (0.8)	Medium (1.2)	High (1.6)
No Skew			
Ratio: 1:1	0.003	0.000	0.000
4:1	0.008	0.000	0.000
Skew			
Ratio: 1:1	0.000	0.000	0.000
4:1	0.000	0.000	0.000

Summary of Results for the OLR

The OLR had excellent Type I Error control across all study conditions; the mean Type I Error was 0.043. The Type I Error of OLR-overall only exceeded 0.05 in five out of twenty-four conditions; it never exceeded 0.10. Furthermore, the Type I Error of the OLR was unrelated to any of the study conditions. The OLR-overall had excellent power for uniform DIF. Power exceeded 0.80 in all conditions and the mean power was high at 0.973. Power for uniform DIF increased with item discrimination and decreased slightly with a 4:1 sample size ratio. Power for

nonuniform DIF was also very high; the mean was 0.934, and only one value was below 0.80. As with GMH, power for nonuniform DIF had a slight U-shaped relationship to item discrimination; power decreased slightly at moderate levels of discrimination, but was highest at high levels. Power for nonuniform DIF was lower with a 4:1 ratio, and when the data were skewed. In terms of discriminating between types of DIF, OLR-uniform performed generally well, although this form of Type I Error was higher than Type I Error for null DIF. As with LDFA-uniform, the Type I Errors were unacceptable only when item discrimination was low or moderate and sample size ratio was 4:1. OLR-nonuniform had very low Type I Errors when uniform DIF was present.

The Impact of Effect Size

One of the goals of the present study was to evaluate the impact of including effect size in DIF detection procedures. In this section, tables showing the difference in Type I Error for each procedure with and without inclusion of a measure of effect size are presented. Tables showing the effect on power of the various procedures with and without effect size are presented separately. In considering power, it is important to note that although DIF was always present in all DIF conditions, the magnitude of nonuniform DIF varied with different study conditions, and particularly, with item discrimination. Thus, the change in power reflects change in DIF magnitude under some circumstances.

The Mantel

Detecting DIF in Test Items

Type I Error. Without effect size, the Type I Error rates for the Mantel procedure when there was no DIF in the data ranged from 0.04 to 0.08 (see Table 29). These rates exceeded 0.05 in thirteen out of twenty-four conditions. The Type I Error rates for the Mantel were substantially reduced when the effect size measure was included; reductions ranged from 0.015 to 0.057 and most were over 0.02. The mean Type I Error for the Mantel across all study conditions was 0.054 without effect size; the mean Type I Error with effect size was 0.022. Thus, the mean decrease in Type I Error of the Mantel procedure with the inclusion of effect size was equal to 0.033; this difference was statistically significant ($z = 11.8$, $p < .001$).

Power for Uniform DIF. The power of the Mantel procedure to detect uniform DIF was extremely high with and without effect size (no table shown). The mean power was 0.994 without effect size and 0.983 with the inclusion of effect size in the decision rule. This mean decrease of 0.011 in power was statistically significant ($z = 5.7$, $p < .01$).

Power for Nonuniform DIF. The power of the Mantel for detecting nonuniform DIF was extremely low in all conditions, regardless of whether or not a measure of effect size was used (no table shown). The mean power of the Mantel procedure to detect uniform DIF was 0.102 without effect size, and 0.096 with the inclusion of effect size. This mean decrease of 0.006 was not statistically significant ($z = 1.4$, $n = 400$, n.s.).

Table 29. Type I Error of the Mantel With and Without Effect Size

Equal Abilities (R-*N* (0,1), F-*N* (0,1)

	Studied Item Discrimination								
	Low (0.8)			Medium (1.2)			High (1.6)		
	Without	**With**	**Diff.**	**Without**	**With**	**Diff.**	**Without**	**With**	**Diff.**
No Skew									
Ratio: 1:1	0.075	0.018	- 0.057	0.040	0.015	- 0.025	0.040	0.010	- 0.030
4:1	0.055	0.020	- 0.035	0.055	0.020	- 0.035	0.043	0.018	- 0.025
Skew									
Ratio 1:1	0.053	0.030	- 0.023	0.048	0.010	- 0.038	0.48	0.028	- 0.020
4:1	0.055	0.020	- 0.035	0.058	0.020	- 0.038	0.40	0.015	- 0.025

Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)

	Studied Item Discrimination								
	Low (0.8)			Medium (1.2)			High (1.6)		
	Without	**With**	**Diff.**	**Without**	**With**	**Diff.**	**Without**	**With**	**Diff.**
No Skew									
Ratio: 1:1	0.053	0.025	- 0.028	0.055	0.015	- 0.040	0.075	0.043	-0.032
4:1	0.053	0.013	- 0.040	0.040	0.025	- 0.015	0.065	0.028	-0.037
Skew									
Ratio: 1:1	0.045	0.018	- 0.027	0.048	0.008	- 0.040	0.080	0.035	- 0.045
4:1	0.068	0.025	- 0.043	0.048	0.013	- 0.025	0.065	0.023	-0.042

The GMH

Detecting DIF in Test Items

Type I Error. The Type I Error of the GMH ranged from 0.033 to 0.07 without the inclusion of effect size and from 0.015 to 0.065 with the inclusion of effect size (see Table 30). Without effect size, Type I Error exceeded 0.05 in 15 out of 24 conditions; with effect size, Type I Error exceeded 0.05 a total of five times. The mean Type I Error was 0.0518 without effect size, and 0.044 with effect size; the difference of 0.0079 was statistically significant ($z = 3.6$, $p < .01$).

Power for Uniform DIF. The GMH had excellent power for uniform DIF, and there was no difference in power when effect size was used (no table shown).

Table 30. The Type I Error of the GMH Procedure With and Without Effect Size

	Equal Abilities (R-*N* (0,1), F-*N* (0,1)								
	Studied Item Discrimination								
	Low (0.8)			Medium (1.2)			High (1.6)		
	With out	**With**	**Diff.**	**With out**	**With**	**Diff.**	**With out**	**With**	**Diff.**
No Skew									
Ratio: 1:1	0.040	0.033	- 0.007	0.048	0.040	- 0.008	0.033	0.030	- 0.003
4:1	0.043	0.033	- 0.012	0.048	0.045	- 0.01	0.058	0.048	- 0.010
Skew									
Ratio 1:1	0.060	0.053	- 0.007	0.055	0.035	- 0.020	0.060	0.053	- 0.007
4:1	0.053	0.048	- 0.005	0.060	0.055	- 0.005	0.053	0.048	- 0.005
	Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)								
	Studied Item Discrimination								
	Low (0.8)			Medium (1.2)			High (1.6)		
	With out	**With**	**Diff.**	**With out**	**With**	**Diff.**	**With out**	**With**	**Diff.**
No Skew									
Ratio: 1:1	0.050	0.038	- 0.012	0.053	0.038	- 0.015	0.070	0.065	- 0.005
4:1	0.053	0.038	- 0.015	0.050	0.045	- 0.005	0.053	0.048	- 0.005
Skew									
Ratio: 1:1	0.048	0.040	- 0.008	0.035	0.023	- 0.012	0.068	0.058	- 0.010
4:1	0.053	0.048	- 0.005	0.048	0.038	- 0.01	0.053	0.050	- 0.003

Note: Difference is With - Without

Power for Nonuniform DIF. The GMH also had very high power for nonuniform DIF, both with and without effect size (Table 31).

Table 31. Power of the GMH Procedure for Nonuniform DIF With and Without Effect Size

	Equal Abilities (R-*N* (0,1), F-*N* (0,1)								
	Studied Item Discrimination								
	Focal 0.8, Ref : 1.8			Focal: 1.2, Ref: 2.5			Focal: 1.6, Ref: 3.2		
	With out	**With**	**Diff.**	**With out**	**With**	**Diff.**	**With out**	**With**	**Diff.**
No Skew									
Ratio: 1:1	1.00	0.985	- 0.015	1.00	0.863	- 0.137	1.00	1.00	0
4:1	1.00	0.953	- 0.047	1.00	0.900	- 0.10	1.00	1.00	0
Skewness									
Ratio 1:1	1.00	0.975	- 0.025	1.00	0.748	- 0.252	1.00	0.99	- 0.01
4:1	1.00	0.955	- 0.045	1.00	0.810	- 0.190	1.00	0.99	- 0.01
	Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)								
	Studied Item Discrimination								
	Focal 0.8, Ref : 1.8			Focal: 1.2, Ref: 2.5			Focal: 1.6, Ref: 3.2		
	With out	**With**	**Diff.**	**With out**	**With**	**Diff.**	**With out**	**With**	**Diff.**
No Skew									
Ratio: 1:1	1.00	0.988	- 0.012	1.00	0.875	- 0.125	1.000	1.000	0
4:1	1.00	0.953	- 0.047	1.00	0.850	- 0.150	0.995	0.993	0.002
Skewness									
Ratio: 1:1	1.00	0.973	- 0.037	1.00	0.850	- 0.150	1.000	0.998	-0.0002
4:1	1.00	0.933	- 0.067	1.00	0.838	- 0.172	1.000	0.990	- 0.01

Without effect size, power was perfect (1.00) in twenty-three of twenty-four conditions. When effect size was included, power was more variable, ranging from 0.748 to 1.00 (see Table 31). The mean power for GMH-overall was 0.999 with effect size and 0.934 without effect size; the difference of 0.066 was statistically significant ($z = 20.85$, $p < .001$).

The LDFA

Detecting DIF in Test Items

Type I Error. Inclusion of the effect size measure had little impact on Type I Error of LDFA; it was 0.0509 without effect size, and 0.0496 with effect size. This difference of 0.0013 was not statistically significant.

Power for Uniform DIF. Inclusion of the effect size measure did not change the power of the LDFA-overall for uniform DIF (not shown). In fact, only two numbers out of twenty-four numbers changed, and these changed very little (not shown). The mean power of LDFA-overall for uniform DIF was 0.989 without effect size, and 0.9866 with effect size.

Power for Nonuniform DIF. As with Type I Error and power for uniform DIF, effect size had very little impact on the power of LDFA to detect nonuniform DIF. Power did decrease slightly in five out of twenty-four situations; these changes were very small (no table shown). The mean power for nonuniform DIF was 0.55 when effect size was not included; when effect size was included the mean power for nonuniform DIF was 0.548.

Differentiation Between Uniform and Nonuniform DIF

The inclusion of effect size led to a significant reduction in the misidentification of nonuniform DIF as uniform DIF for the LDFA. Without effect size, the Type I Error rates for LDFA-uniform when nonuniform DIF was present ranged from 0.030 to 0.243; with effect size, the Type I Errors ranged from 0.03 to 0.173 (see Table 32). Interestingly, effect size had a large impact when sample size ratio was 4:1 but made little difference when sample size ratio was 1:1 (see Table 32).

Table 32. The Type I Error of the LDFA-uniform When Nonuniform DIF was Present With and Without Effect Size

	Equal Abilities (R-*N* (0,1), F-*N* (0,1)								
	Studied Item Discrimination								
	Focal 0.8, Ref : 1.8			Focal: 1.2, Ref: 2.5			Focal: 1.6, Ref: 3.2		
	With out	**With**	**Diff.**	**With out**	**With**	**Diff.**	**With out**	**With**	**Diff.**
No Skew									
Ratio: 1:1	0.055	0.055	0	0.040	0.040	0	0.043	0.043	0
4:1	0.113	0.043	- 0.07	0.120	0.073	- 0.053	0.083	0.048	- 0.035
Skew									
Ratio 1:1	0.050	0.050	0	0.043	0.043	0	0.053	0.053	0
4:1	0.138	0.065	- 0.063	0.088	0.043	- 0.045	0.093	0.055	- 0.038
	Unequal Abilities (R-*N* (0,1), F-*N* (-0.5,1)								
	Studied Item Discrimination								
	Focal 0.8, Ref : 1.8			Focal: 1.2, Ref: 2.5			Focal: 1.6, Ref: 3.2		
	With out	**With**	**Diff.**	**With out**	**With**	**Diff.**	**With out**	**With**	**Diff.**
No Skew									
Ratio: 1:1	0.173	0.173	0	0.058	0.058	0	0.030	0.030	0
4:1	0.243	0.170	- 0.073	0.098	0.068	- 0.03	0.090	0.058	- 0.042
Skew									
Ratio: 1:1	0.153	0.153	0	0.073	0.073	0	0.058	0.058	0
4:1	0.223	0.138	- 0.085	0.123	0.070	- 0.053	0.093	0.068	- 0.025

When the sample size ratio was 1:1 there was no reduction in Type I Error with the use of the effect size measure. However, when the sample size ratio was 4:1, the mean Type I Error was reduced from 0.125 to 0.075 with the inclusion of effect size. Overall, the mean reduction in Type I Error for differentiating between uniform and nonuniform DIF was 0.0251; this was statistically significant ($z = 8.6$, $p < .001$).

The Ordinal Logistic Regression Procedure

Identifying DIF in Test Items

Type I Error. The Type I Error of the OLR-overall was low both with and without effect size. As shown in Table 33, inclusion of the effect size measure reduced the Type I Error of the OLR-overall slightly in twenty out of twenty-four study conditions. These reductions ranged from 0.003 to 0.013; most were 0.005. The mean reduction in Type I Error of the OLR was 0.0051; this was nonsignificant ($z = 1.2$, $p < .23$).

Power for Uniform DIF. Inclusion of the effect size measure did not affect the power of OLR-overall to detect uniform DIF; power was high ($\underline{M} = 0.96$) with and without effect size across all twenty-four conditions (no tables shown).

Table 33. Type I Error of the OLR Procedure With and Without Effect Size

Equal Abilities (R-*N* (0,1), F-*N* (0,1)

	Studied Item Discrimination								
	Low (0.8)			Medium (1.2)			High (1.6)		
	With out	**With**	**Diff.**	**With out**	**With**	**Diff.**	**With out**	**With**	**Diff.**
No Skew									
Ratio: 1:1	0.048	0.038	- 0.010	0.040	0.035	- 0.005	0.03	0.025	- 0.005
4:1	0.053	0.050	- 0.003	0.053	0.048	- 0.005	0.05	0.043	- 0.005
Skew									
Ratio 1:1	0.048	0.043	- 0.005	0.053	0.04	- 0.013	0.035	0.025	- 0.010
4:1	0.053	0.050	- 0.003	0.048	0.045	- 0.003	0.043	0.043	0

Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)

	Studied Item Discrimination								
	Low (0.8)			Medium (1.2)			High (1.6)		
	With out	**With**	**Diff.**	**With out**	**With**	**Diff.**	**With out**	**With**	**Diff.**
No Skew									
Ratio: 1:1	0.075	0.068	- 0.007	0.050	0.043	- 0.007	0.065	0.055	- 0.010
4:1	0.025	0.025	0	0.045	0.040	- 0.005	0.05	0.045	- 0.005
Skew									
Ratio: 1:1	0.058	0.058	0	0.058	0.053	- 0.005	0.045	0.040	- 0.005
4:1	0.055	0.055	0	0.038	0.033	- 0.005	0.048	0.043	- 0.005

Power for Nonuniform DIF. As shown in Table 34, inclusion of the effect size measure did affect the power of OLR-overall to detect nonuniform DIF. The mean power for nonuniform DIF was 1.00 without effect size, and was 0.934 with effect size included. This difference of 0.066 was statistically significant ($z = 20.85, p < .001$).

Table 34. Power of the OLR-overall for Nonuniform DIF With and Without Effect Size

Equal Abilities (R-*N* (0,1), F-*N* (0,1)

	Studied Item Discrimination								
	Focal 0.8, Ref : 1.8			Focal: 1.2, Ref: 2.5			Focal: 1.6, Ref: 3.2		
	Without	**With**	**Diff.**	**Without**	**With**	**Diff.**	**Without**	**With**	**Diff.**
No Skew									
Ratio: 1:1	1.00	0.985	- 0.015	1.00	0.863	- 0.137	1.00	1.00	0
4:1	1.00	0.963	- 0.047	1.00	0.900	- 0.10	1.00	1.00	0
Skew									
Ratio 1:1	1.00	0.975	- 0.025	1.00	0.748	- 0.252	1.00	0.99	- 0.01
4:1	1.00	0.955	- 0.045	1.00	0.810	- 0.190	1.00	0.99	- 0.01

Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)

	Studied Item Discrimination								
	Focal 0.8, Ref : 1.8			Focal: 1.2, Ref: 2.5			Focal: 1.6, Ref: 3.2		
	Without	**With**	**Diff.**	**Without**	**With**	**Diff.**	**Without**	**With**	**Diff.**
No Skew									
Ratio: 1:1	1.00	0.982	- 0.018	1.00	0.875	- 0.125	1.00	1.00	0
4:1	1.00	0.953	- 0.047	1.00	0.850	- 0.150	1.00	0.998	- 0.002
Skew									
Ratio: 1:1	1.00	0.963	- 0.037	1.00	0.850	- 0.150	1.00	0.998	- 0 .002
4:1	1.00	0.933	- 0.067	1.00	0.838	- 0.172	1.00	0.99	- 0.01

Discrimination Between Uniform and Nonuniform DIF

Inclusion of the effect size measure reduced Type I Error of OLR-uniform when nonuniform DIF was present in the data. Without effect size, Type I Error ranged from 0.003 to

0.333. When effect size was included in the decision rule, the Type I Error of OLR-uniform for differentiating between types of DIF was reduced in twenty-three out of twenty-four conditions (see Table 35).

Table 35. The Type I Error for OLR-uniform When only Nonuniform DIF was Present

Equal Abilities (R-*N* (0,1), F-*N* (0,1)

Studied Item Discrimination

	Focal 0.8, Ref : 1.8			Focal: 1.2, Ref: 2.5			Focal: 1.6, Ref: 3.2		
	Without	**With**	**Diff.**	**Without**	**With**	**Diff.**	**Without**	**With**	**Diff.**
No Skew									
Ratio: 1:1	0.025	0.008	- 0.017	0.015	0.013	- 0.002	0.033	0.018	- 0.015
4:1	0.183	0.105	- 0.078	0.123	0.065	- 0.058	0.148	0.060	- 0.088
Skew									
Ratio 1:1	0.003	0.003	0	0.003	0.00	- 0.003	0.008	0.003	- 0.005
4:1	0.155	0.100	- 0.055	0.083	0.030	- 0.053	0.135	0.058	- 0.077

Unequal Abilities (R-*N* (0,1), F-*N* (- 0.5,1)

Studied Item Discrimination

	Focal 0.8, Ref : 1.8			Focal: 1.2, Ref: 2.5			Focal: 1.6, Ref: 3.2		
	Without	**With**	**Diff.**	**Without**	**With**	**Diff.**	**Without**	**With**	**Diff.**
No Skew									
Ratio: 1:1	0.070	0.055	- 0.015	0.010	0.003	- 0.007	0.018	0.003	- 0.015
4:1	0.333	0.303	- 0.030	0.190	0.125	- 0.065	0.183	0.095	- 0.088
Skew									
Ratio: 1:1	0.058	0.053	- 0.005	0.013	0.010	- 0.003	0.018	0.008	- 0.010
4:1	0.250	0.228	- 0.022	0.130	0.115	- 0.015	0.100	0.078	- 0.022

With effect size included, the Type I Error ranged from 0.003 to 0.303. Reductions ranged from 0.003 to 0.088; most differences were larger than 0.020. The mean Type I Error of the OLR-uniform test when nonuniform DIF was present was 0.096 without effect size; this was reduced to a mean of 0.064 with the inclusion of effect size. This difference of 0.031 was statistically significant ($z = 3.57$, $p < .001$). The Type I Error of OLR-nonuniform when uniform DIF was present was extremely low and was not changed by the inclusion of effect size.

Summary of Findings on Power and Type I Error

A summary of results for all four procedures can be found in Table 36. It shows: Type I Error for all procedures as well as the number of times (out of 24) that Type I Error exceeded .05, power for uniform DIF and the number of times that power was below .8, power for nonuniform DIF as well as the number of times that power was below .8, and Type I Error for discriminating types of DIF. As shown, all procedures had very low Type I Error rates, but the Mantel had the lowest rate. Similarly, all had very high, and nearly equivalent power for uniform DIF. However, the procedures differed widely in their power to detect nonuniform DIF: the Mantel was unable to detect uniform DIF, while the LDFA had low power for detecting nonuniform DIF. On the other hand, the GMH and OLR had excellent power for nonuniform DIF. The LDFA and the OLR were the only procedures that were capable of differentiating between uniform and nonuniform DIF; the OLR had slightly lower Type I Error for misidentifying nonuniform DIF as uniform DIF.

Table 36. Summary of results for the four procedures

Procedure	Type I Error. Times (out of 24) above .05	Power for Uniform DIF. Times below .8	Power for Nonuniform DIF. Times below .8	Type I Error: Nonuniform identified as uniform
Mantel	0.021 0 times	0.98 0 times	0102 24 times	n.a.
GMH	0.044 4 times	0.98 0 times	0.934 1 time	n.a.
OLR	0.044 4 times	0.97 0 times	0.934 1 time	0.064
LDFA	0.0496 13 times	0.99 0 times	0.55 18 times	0.072

Summary of Findings for Effect Size

The analyses of the impact of including effect size produced somewhat mixed results. Inclusion of effect size resulted in decreased Type I Error for all procedures; reductions in Type I Error of the Mantel were substantial ($\underline{M}$ = 0.033). Inclusion of effect size also resulted in significant reductions in Type I Error for the GMH ($\underline{M}$ = 0.0079) and marginally significant results for the OLR-overall ($\underline{M}$ = 0.0056); effect size did not change the Type I Error of the LDFA-overall. Inclusion of effect size led to substantial decreases in Type I Error of both the LDFA-uniform ($\underline{M}$ = 0.025) and OLR-uniform ($\underline{M}$ = 0.031) for discriminating between uniform and nonuniform DIF.

Inclusion of the effect size measure did not affect the power of the GMH, LDFA,or OLR procedures to detect uniform DIF; it did however, result in a small, but significant decrease in the power of the Mantel to detect uniform DIF. The inclusion of effect size did not have any impact on power of the Mantel or LDFA to detect nonuniform DIF. However, inclusion of effect size did result in moderate decreases in power to detect nonuniform DIF for the OLR-overall and the GMH.

Summary of findings by hypothesis

Table 37 presents a summary of the study hypotheses, and whether they were met, not met, or partially met. Most of the study hypotheses were met or partially met. For example, the hypothesis that all procedures would have very low Type I Error was met while the hypothesis that the OLR would have the lowest Type I Error was not met. The hypothesis that the Mantel would have the highest power for uniform DIF was not met while the hypothesis that it would have no power for nonuniform DIF was met. The hypothesis that the GMH would have high power for uniform and nonuniform DIF was met. Finally, the hypothesis that the OLR and LDFA would both have excellent power for uniform and nonuniform DIF was partially met; the LDFA actually had rather low power for nonuniform DIF.

Table 37. Results by hypothesis

Hypothesis	Met	Partially Met	Not Met	Comments
1.1 Type I Error all low	√			
1.2 OLR lowest			√	
2.1 Mantel highest for uniform DIF			√	
2.2 Mantel unable to detect nonuniform	√			
2.3 GMH high for uniform/ nonuniform	√			
2.4 OLR and LDFA high for both		√		
2.5 OLR and LDFA highest for nonuniform		√		LDFA low
3.1 High Abil * High Disc = Type I Error		√		small effect
3.2 Ratio of 4:1 lower power	√			
4.1 Effect size lower Type I Error		√		
4.2 Effect size lower power	√			

CHAPTER V
DISCUSSION

Bias arises when educational test items are asked in a way that prevents certain groups of examinees from demonstrating their knowledge; it thus results in unfair testing (Camilli & Shepard, 1994). The purpose of DIF detection procedures is to identify potentially biased items, and thereby, ultimately, to increase the validity and fairness of tests. Systematic application of methods for detecting DIF, corrective action, and subsequent demonstrations that inferences made from tests are unbiased are vital steps in test validation and will help to improve equity in testing (Camilli & Shephard, 1994). "The development and evaluation of DIF procedures for detecting potentially biased items is a critical area of research for psychometricians " (Hambleton, Clauser, Mazor & Jones, 1993, p. 1). In this study, evidence on the validity and practicality of four DIF detection procedures was presented. It represents an addition to the growing body of literature on DIF detection. It has also made several original contributions to the literature:

1. This is the first empirical study of Ordinal Logistic Regression. As might be expected with a new technique, some refinements were necessary and these were made. OLR was expected to perform very well, and it did. This means that professionals who wish to perform DIF detection in applied settings now have another technique at their disposal as well as additional guidelines (see Conclusions and Recommendations) on how and when to use them.

2. Practical significance is concerned with whether or not results are meaningful in the real world (Kirk, 1996). Although many experts are now convinced of the importance of assessing practical significance as well as statistical significance, little work has been done on incorporating measures of effect size into DIF modelling (Zwick, 1996). It is essential to understand how effect size

measures work and to propose guidelines as to how it should be used in decision making. This is the first empirical study to incorporate effect size and significance testing together in DIF modelling and to actually assess the impact of this type of modelling on decision making.

3. The effect size measure used in this study is new, but closely related to several of the leading measures of effect size in DIF. It may prove to be quite useful in applied settings, as it is intuitively reasonable and comprehensible. It does show promise and should be tested further under different circumstances; refinement may also be in order.

4. The effect of skewness, high reference/focal group sample size ratio, and high item discrimination had not been previously well studied; the effects of some of these factors proved to be quite intriguing. This has helped to further understanding of the strengths and limitations of these techniques and also of when they may and may not be used with real data.

Strengths and Limitations of this Study

Strengths

This study has several strengths:

1. This study was carefully conceived and developed. The validity of the data generation program was carefully assessed, as was the validity of each program used to run the DIF assessment procedures. All programs for data generation and DIF detection were carefully pretested and evaluated to confirm that 1) the programs were producing valid results and 2) the data met the required assumptions. This careful validation means that one can have confidence in study results. Furthermore, without this pretesting, the fact that the study data did not meet assumptions would not have been discovered, and OLR would have been inappropriately applied to this data. Also,

the LDFA would have had extremely high Type I Errors; simply adding cubic and quadratic terms to the model reduced this to very reasonable levels.

2. The use of 400 replications meant that confidence intervals around estimates are quite small and that estimates are more precise than they would have been with fewer replications.

3. The four DIF detection procedures were evaluated under a variety of realistic testing conditions.

4. Using the measures of effect size in the study led to a very clear understanding of the data and of how the data generation process worked

5. No-one has used the type of sequential modelling recommended by Zumbo (1999) and Swaminathan and Rogers (1990) in an empirical test of DIF assessment procedures. Furthermore, no other researcher has studied the Type I Error of the procedures in discriminating between uniform and nonuniform DIF.

Limitations

This study also has some limitations:

1. Simulation studies are essential for studying the Type I Error and power of DIF detection procedures because the data has known parameters. However, generalizability is limited to the conditions under study.

4. The sample size used in this study was very large (N = 4000). In many testing situations, sample size may be much smaller. Holding sample size constant did not allow for study of its effect on power.

3. In this study, DIF was limited to one item. This is unlikely to happen in real test situations. When DIF is present in several items, it becomes difficult to detect because the matching test is more contaminated.

4. Only one type of DIF was simulated at one time in an item in this study. In real test situations, an item may include both uniform and nonuniform DIF.

5. In this study only two levels of group ability difference were examined: no group ability difference and a small (0.5) group ability difference. This means that we do not yet know how OLR performs when group ability differences are larger.

Despite some limitations, the study should provide valuable information on the validity and utility of the Mantel, GMH, LDFA, and OLR procedures. In the following sections, study results are discussed and interpreted.

Type I Error

The hypothesis that the four DIF detection procedures would have low Type I Error rates was upheld; all four procedures had good control over Type I Error under all combinations of sample size ratio, item discrimination, group ability difference, and skewness. Most rates were below 0.05; none exceeded 0.10. The procedures were nearly indistinguishable in terms of Type I Error. In a comparison of the four main tests (Mantel, GMH, LDFA-overall, and OLR-overall), the Mantel had relatively the lowest Type I Error rate at 0.021, and the LDFA-overall relatively the highest at 0.0496; the GMH and OLR had equivalent rates of 0.044. These rates were all quite acceptable, although the Type I Error rates for the LDFA-overall were slightly more problematic than those for the GMH and OLR-overall. For example, the Type I Error rates for LDFA-overall

exceeded 0.05 in thirteen out of twenty-four conditions while the Type I Error rates for the GMH and OLR-overall exceeded 0.05 only four and five times (respectively) out of twenty-four and those for the Mantel (with effect size) never exceeded 0.05.

The Mantel is only designed to detect one type of DIF while the other three procedures are designed to detect both uniform and nonuniform DIF; in order to detect both types of DIF, the Mantel would have to be paired with a test that could detect nonuniform DIF. It might therefore be more realistic to compare the Mantel with the components of the other procedures that were designed to detect only one type of DIF. This comparison revealed no difference between the procedures (Mantel ($\underline{M}$ = 0.021, LDFA-uniform ($\underline{M}$ =0.023), LDFA-nonuniform ($\underline{M}$ =0.026), OLR-uniform ($\underline{M}$ =0.024), and OLR-nonuniform ($\underline{M}$ =0.023)).

The tests for one type of DIF only all had slightly lower Type I Error rates than the tests for both types of DIF (GMH ($\underline{M}$ = 0.044), LDFA-overall ($\underline{M}$ = 0.0496), and OLR-overall ($\underline{M}$ = 0.044)). This is not surprising; the tests which detect both uniform and nonuniform DIF are designed to detect differences between groups in any direction and will be more inclusive than those which are designed to detect only one type of DIF. Therefore, they should naturally pick up more noise than the more limited tests. Because it is important to detect both types of DIF, the Type I Error rates for the tests which detect both types of DIF are probably more representative of reality.

The finding that the four procedures had very similar Type I Error rates is consistent with results from other studies. For example, Chang et al. (1996), Spray and Miller (1994), Tian (1999), and Zwick et al. (1997) all report that polytomous DIF detection procedures had very low

and nearly indistinguishable Type I Error rates when no group ability difference existed and in some cases when there were differences in group ability (Chang et al.,1996; Zwick et al., 1997).

There are some similarities and some differences between the magnitude of Type I Errors found in this study and those from other studies. For example, the mean Type I Error for the Mantel without effect size (0.054) is very close to the mean of 0.049 reported by Chang et al.. (1996) in their Study 1. The differences between this and other studies may be largely explained by differences in study design, including the number of replications, reliability of the matching test, and the number and type of independent variables. For example, the Type I Error rates for the Mantel, GMH, and LDFA reported by Spray and Miller (1993) were slightly lower than those in the present study. However, they had a very simple study design; item discrimination was low and there were no group ability differences.

The Type I Error rates for the Mantel, GMH, LDFA, and OLR were not related to skewness, sample size ratio, or to the main effects of item discrimination and group ability difference. However, for the Mantel, GMH, and LDFA there was a significant interaction between group ability difference and studied item discrimination. Type I Error rates were slightly higher when studied item discrimination was high and when there was a small (- 0.5) group ability distribution difference. The Type I Error rates for the OLR procedure were unaffected by any of the studied factors. This is ideal; it was hoped that the study conditions would not affect the rejection rates for items with no DIF.

In this study, group ability distribution did not have a main effect on the Type I Error rates of the Mantel, GMH, LDFA, or OLR. This finding is consistent with findings by Chang et al.,

(1996), who reported that large differences in group ability did not affect Type I Error rates for the Mantel and STND (Study 1). It is also consistent with findings by Zwick et al. (1997), who also reported that large differences in group ability had little effect on the STND. However, they are inconsistent with the results reported by Tian (1999) who found highly inflated Type I Error rates with even moderate group ability differences. The reasons for this inconsistency will be discussed below.

The finding of higher Type I Error rates for the Mantel, GMH, and LDFA procedures under conditions of group ability differences and high studied item discrimination is consistent with results from Chang et al. (1996), Tian (1999), and Zwick et al.(1997) who all reported inflated Type I Error rates with group ability differences and high studied item discrimination. In the present study, the effect of ability difference and item discrimination was small and relatively unimportant; mean rates under conditions of group ability differences and high item discrimination ranged from 0.025 to 0.088. These rates are lower than the high Type I Error rates (e.g, 0.35, 0.41) reported by Chang et al. (1996) and Zwick et al. (1997) for conditions of high studied item discrimination and large (1.0 difference) group ability differences. However, this difference is easily explained by the fact that in the present study, only moderate (0.5) differences in group ability distribution were simulated. It seems that the synergistic effect of item discrimination and group ability difference is much stronger when group ability differences are large than when group ability differences are small.

Tian (1999) is the only other researcher who has also studied moderate (0.5) differences in group ability. Interestingly, the mean Type I Error rates for the Mantel, LDFA, and GMH for the condition of high studied item discrimination and moderate group ability differences are much

lower in this study than in hers. In fact, mean Type I Error rates in the present study are generally lower than those found in her study (although they are more consistent with those reported by Chang et al.(1996) and Zwick et al. (1997). These differences are probably largely due to a difference in the reliability of the matching test; they may also be due to a difference in the number of replications. The matching test used in the present study consisted of 26 polytomous items; the total test score ranged from 0 to 78. On the other hand, the matching tests used by Tian (1999) consisted largely of dichotomous items; 20% were polytomous. The total score in her tests ranged from 0 to 25 for the 20 item test and from 0 to 61 for the 40 items. Type I Error and power are both directly related to test length because in general, longer tests are more reliable than shorter tests (Streiner & Norman, 1995). For example, Tian found that Type I Error rates were much lower for her 40 item test than for the 20 item test; power was also higher for the 40 item test. Thus, the lower overall Type I Errors in this study may be partially explained by the fact that the score range in the test used in the present study was wider (and the score possibly more reliable) than that for either of Tian's tests. This hypothesis is supported by the fact that Type I Error rates from Tian's 40 item test were much more similar to those in the present study than Type I Error rates from her 20 item test. Another difference between this study and that performed by Tian (1999) is that Tian performed 100 replications while the present study involved 400 replications. Therefore, the confidence intervals around estimates of Type I Error would be wider in her study. In fact, the confidence intervals around many of the Type I Error values for the 40 item test in her study would include values found in the present study.

Summary: Type I Error

In summary, all four procedures performed well in terms of Type I Error rates under the conditions used in the present study. The Mantel had the lowest mean rates, but the differences among the procedures was very small (0.02). Furthermore, because the Mantel only detects uniform DIF, it may be more realistic to compare the Mantel to the components of the LDFA and OLR which detect uniform DIF. This comparison showed no differences between the Mantel, OLR, and LDFA.

The OLR seems to be less affected by the combination of group ability difference and high item discrimination than the other DIF assessment procedures. However, for all procedures, the effect of the independent variables was small, and relatively unimportant; rates remained well below 0.10. All four could be recommended on the basis of their Type I Error performance, although the fact that Type I Error exceeded 0.05 in thirteen conditions raises some concerns.

Because Type I Error was well under control, power comparisons can be legitimately made for all four procedures.

Uniform DIF

The performance of the Mantel, GMH, LDFA, and OLR in detecting uniform DIF in polytomous items was consistently excellent. As hypothesized, all four procedures had extremely high power for detecting uniform DIF; the mean power for uniform DIF was 0.98, 0.98, 0.99, and 0.963 for the Mantel, GMH, LDFA, and OLR procedures, respectively. Although power rates varied slightly across different conditions, all values remained well above the minimum acceptable value of 0.80. The finding that all procedures had excellent and nearly equivalent power for

uniform DIF is consistent with existing literature. For example, Spray and Miller (1994) reported that the Mantel, GMH, and LDFA-uniform all had perfect (1.00) power to detect uniform DIF when the total sample size was 4000. Tian (1999) found that when the total sample size was 2400, the Mantel, GMH, and LDFA-uniform all had power for uniform DIF that ranged from 0.88 to 0.95 with low studied item discrimination, from 0.99 to 1.00 with moderate discrimination, and perfect power with high studied item discrimination. These rates are almost identical to those from the present study; the finding that LDFA-uniform had relatively the highest power for uniform DIF is also consistent. The findings of extremely high power for uniform DIF would also be expected with such a large sample size, because power increases markedly with increased sample size. The fact that Type I Error rates were low under all conditions suggests that these power results are valid. Polytomous DIF detection procedures may have excellent power even with smaller (N = 1000) sample sizes, if item discrimination is moderate or high (Chang et al.,1996; Zwick et al., 1997). Thus, it seems that most DIF detection procedures work quite well in identifying uniform DIF in test items.

The power of the LDFA procedure to detect uniform DIF changed very little across the different study conditions. The power of the Mantel, GMH, and OLR for uniform DIF was related to sample size ratio and item discrimination; their power for uniform DIF was slightly higher with moderate and high studied item discrimination. Tian (1999) reported similar findings for sample size ratio and for item discrimination. In her study, power for uniform DIF was somewhat lower when group sample sizes were unequal (2:1) than when they were equal. It is not surprising that unequal sample sizes can result in slightly lower power; large differences in group sample sizes mean that, at each level of total score, there are relatively fewer examinees in the focal group to

compare to those in the reference group. Thus, the effective sample size is smaller and power is lower.

The finding that power for uniform DIF increased with increased item discrimination was also expected; Chang et al. (1996), Tian (1999), and Zwick et al. (1997) have all found that most procedures have very high power for uniform DIF when the studied items have moderate and high item discrimination. The fact that power increased can be explained by the fact that the actual magnitude of uniform DIF increases with high item discrimination (Chang et al., 1996; Zwick et al., 1997).

Summary: Power for Uniform DIF

In summary, all procedures had extremely high power for uniform DIF. They were almost indistinguishable; the LDFA had relatively the highest mean power; its power was only 0.03 higher than that of the OLR, which had the lowest mean power of 0.96. Although several factors were related to the power of these procedures, their effects were not very important; in all cases, power for uniform DIF under all conditions was well over 0.80. Thus, these procedures could all be used to detect uniform DIF under the conditions studied here.

Power for Nonuniform DIF

Prior to the study, it was hypothesized that the GMH, LDFA, and OLR would all be able to detect nonuniform DIF, and that the Mantel would have little or no power to do so. These hypotheses were largely upheld, although power results for LDFA were somewhat disappointing.

The Mantel was unable to detect nonuniform DIF. This was expected, as it was not designed to do so. The Mantel is a signed test, and thus is sensitive to mean differences (see

Equation 5); differences in opposite directions (as in nonuniform and balanced DIF) cancel each other out (Zwick et al., 1993). Spray and Miller (1994) and Tian (1999) also report that the Mantel has extremely low power for nonuniform DIF while Ankenmann et al. (1999), Chang et al. (1996), Tian (1999), and Zwick et al. (1993) all report that the Mantel cannot detect balanced DIF.

The GMH and OLR had excellent power for detecting nonuniform DIF across all study conditions. For both, power was perfect when effect size was not included in the decision rule. With effect size included, power ranged from 0.748 to 1.00; only one value out of twenty-four was unacceptable (below 0.80). The finding that the GMH had very high power for nonuniform DIF was expected on the basis of the way it is calculated. It can be seen from Equation 9 that the GMH is an unsigned test, and should be sensitive to between group differences in the frequencies of any item scores. The effect size measure used with the GMH is also unsigned; regression predicted differences in item score in either direction are calculated. The finding that the OLR had excellent power for nonuniform DIF was also expected; in OLR, nonuniform DIF is specifically modelled in a way that closely matches the definition of nonuniform DIF (an interaction between total score and group membership) and the effect size measure detects differences in either direction.

Several factors were related to the power of the GMH and OLR for detecting nonuniform DIF. These included item discrimination, sample size ratio, skewness, and interactions between item discrimination and sample size ratio and between skewness and ability distribution. However, with the exception of item discrimination, the impact of these study conditions was small. The effect of item discrimination was stronger; power for the GMH and OLR ranged from 0.95 to 0.99 with low item discrimination, dropped to means ranging from 0.75 to 0.88 with moderate discrimination, and increased to 0.99 and 1.00 when item discrimination was high. The

fact that this decrease was only observed when effect size was included in the decision rules for the OLR and GMH is a reflection of the fact that DIF magnitude in the data actually fell below this very conservative threshold (0.03 difference out of 3 points) at times under some study conditions. Although it is difficult to explain why this happened, it is a reflection of actual DIF magnitude rather than a problem with the procedures.

Non-uniform DIF is a very complex phenomenon; its magnitude is related to both the difference in studied item discrimination between the focal and reference group and to the base level of item discrimination in the focal group. For example, in the process of setting up the data generation program, where, as the base level of item discrimination increased, it was necessary to increase the difference between the a-parameters by larger and larger amounts in order to maintain a constant level of nonuniform DIF. Thus, there seems to be a ceiling effect; as item discrimination values become higher, the total amount of nonuniform DIF may be attenuated. The dip in the magnitude of nonuniform DIF when item discrimination in the focal group means that although the aim of increasing a-parameter differences was to maintain a consistent level of nonuniform DIF in the data, the actual magnitude was not quite even across the three levels of item discrimination. Finally, the power results from the OLR-overall and the GMH are actually quite encouraging because these techniques were able to detect very small amounts of nonuniform DIF with very high accuracy.

Power results for LDFA-nonuniform and LDFA-overall for nonuniform DIF were generally disappointing whether modelling was done with or without effect size. LDFA-overall and LDFA-nonuniform both had reasonable power when the studied item discrimination was low, but power decreased to unacceptable levels when focal group item discrimination was moderate or

high. Even at low levels of item discrimination, the LDFA mean of 0.876 was much lower than that of the GMH or OLR (mean 0.964). This finding is contrary to the results of Spray and Miller (1994) who found that LDFA-nonuniform had perfect power (1.00) for nonuniform DIF with a sample size of 4000. However, the studied item discrimination value of 0.5 (reference group = 1.0) was slightly lower than the lowest value used in the present study (0.8). Because the power of LDFA decreases with increasing discrimination, it is not surprising that it was slightly lower in the present study than in that of Spray and Miller (1993). Tian's (1999) findings on th LDFA are more consistent with those of the present study. She reported power values ranging from 0.77 to 0.98 when group abilities were equal and sample size was similar; these values are very similar to those in the present study for low item discrimination and equal group abilities.

The finding that LDFA was unable to detect nonuniform DIF at moderate and high levels of studied item discrimination (1.2 and 1.6 in the focal group; 2.5 and 3.2 in the reference group) is new and was unexpected. However, the LDFA has not yet been well studied, and no-one had previously studied the effect of high item discrimination on the detection of nonuniform DIF. Tian (1999) did study the effect of differences in group ability on power for nonuniform DIF. Her findings are highly relevant to those from the present study as she reported that the power of LDFA-nonuniform for nonuniform DIF decreased steadily as group ability differences increased; power ranged from 0.77 to 0.98 with no ability difference, from 0.59 to 0.84 with a moderate (0.5) ability difference and from 0.19 to 0.61 when there was a large (1.0) group ability difference. Thus, it seems that the use of LDFA may be problematic when item discrimination in the focal group is moderate or high as well as when group ability differences exist. It is difficult to explain why this problem of low power for the LDFA under conditions of high item discrimination

(present study) or group ability difference (Tian, 1999) occurs. This low power is not due to the inclusion of an effect size measure; the LDFA had low power regardless of whether effect size was used or not. Furthermore, the OLR and GMH do not have this problem. They had consistently excellent power without effect size, and very high power with effect size. It seems likely that the low power for nonuniform DIF displayed by the LDFA under certain conditions may be related to the way in which DIF is modelled in LDFA. The OLR and LDFA procedures are closely related; both are based on logistic regression. The main difference between the two is that in the OLR and GMH procedures, the item score is the dependent variable, while in LDFA, group membership is the dependent variable and item score is included among the predictors; nonuniform DIF is modelled as the interaction between item score and total score (see Equation 23, page 17). This does not match the well accepted definition of nonuniform DIF as the interaction between group membership and total score. Although the reversal of dependent variable and predictors seems to work fairly well for uniform DIF, it is problematic for nonuniform DIF. The performance of the LDFA in detecting nonuniform DIF should be studied further in order to learn whether other study conditions affect it.

Summary: Power for Detecting Nonuniform DIF

As expected, the Mantel could not detect nonuniform DIF in test items. The GMH and OLR both had excellent power for nonuniform DIF although this power was somewhat lower when effect size was used. On the other hand, the LDFA-overall had lower power than the OLR and GMH under all conditions; power deteriorated markedly when the studied item discrimination was high. The power of the OLR and GMH was related to several of the study conditions, including item discrimination, sample size ratio, skewness, and the interaction between item

discrimination and sample size ratio. However, with the exception of studied item discrimination, all of these effects were relatively small. Even then, the effect of item discrimination on power was much smaller for GMH and OLR than for the LDFA. The fact that the OLR and GMH had very high power for nonuniform DIF and were relatively robust under various study conditions means that, in general, it would be preferable to use OLR or GMH rather than LDFA. LDFA could be used when studied item discrimination was low.

Differentiation between Uniform and Nonuniform DIF

Differentiation between uniform and nonuniform DIF is the second phase in DIF detection; this classification is important because it will help test developers and researchers understand exactly how the DIF in the item affects people in the two groups according to their ability level. Such knowledge may be useful in decision making about whether to review and, subsequently, whether to retain items.

Of the four procedures evaluated in this study, only the LDFA and OLR could be used to differentiate between uniform and nonuniform DIF. The Mantel and GMH both have only one test; the Mantel can only detect uniform DIF, and the GMH can detect both, but cannot differentiate between them. On the other hand, both the OLR and LDFA have separate tests for uniform and nonuniform DIF. This means that not only should they be able to identify both types of DIF, they should be able to differentiate between them.

The performance of LDFA-uniform for discriminating between uniform and nonuniform DIF was quite good, except when there were group ability differences and item discrimination was low. When there were no group ability differences, the Type I Error of LDFA-nonuniform was

ranged from 0.043 to 0.065. With group ability differences, the Type I Error was well under 0.10, except when item discrimination was low; in this case it ranged from 0.14 to 0.17. The overall mean rate for misidentifying nonuniform DIF as uniform DIF was 0.072. The overall mean rate for the OLR-uniform was somewhat lower, at 0.064. However, the range for the OLR-uniform was much wider (0 to 0.30). While the LDFA-uniform was most affected by the interaction between item discrimination and group ability, the OLR-uniform was most affected by sample size ratio. The Type I Error rates for misidentification of nonuniform DIF as uniform DIF were very low and acceptable (ranging from 0 to 0.055), except when the sample size ratio was 4:1 (ranging from 0.03 to 0.30).

In contrast to the somewhat inflated Type I Error rates for the LDFA-uniform and OLR-uniform, the Type I Error rates for the LDFA-nonuniform and OLR-nonuniform when uniform DIF was simulated were extremely low. Thus, it seems that while nonuniform DIF may be mistakenly identified as uniform DIF, it is extremely unlikely that uniform DIF will be misidentified as nonuniform DIF.

Summary: Classification of Uniform and Nonuniform DIF

In summary, the differentiation between uniform and nonuniform DIF seems to be slightly more difficult than the identification of DIF (or null DIF). This is consistent with the observation note that uniform and nonuniform DIF are not always clearly separable in polytomous items (Swaminathan & Rogers, 1990). Nevertheless, these procedures worked quite well for differentiating between uniform and nonuniform DIF in most study conditions. Rates for the OLR and LDFA were comparable, although the OLR-uniform had a slightly lower overall mean than the

LDFA-uniform. Similarly, the OLR-nonuniform had a slightly lower overall mean Type I Error for the misidentification of uniform DIF as nonuniform DIF. Interestingly, it was extremely rare for uniform DIF to be classified as nonuniform DIF; misclassification of nonuniform DIF is much more probable.

On the basis of these results, both OLR and LDFA procedures could be used for the classification of DIF as predominantly uniform or nonuniform. However, because the LDFA has lower power for nonuniform DIF, it would be best to use the OLR-overall test followed by the OLR-uniform test to first identify and then classify DIF. It would also be wise to include the measure of effect size and to look at the data graphically in order to reduce Type I Error.

The Effect of the Independent Variables

Ideally, the performance of DIF detection procedures would be unaffected by any variation in study conditions. This would enable those who use DIF detection procedures to apply them in a variety of settings. In the present study, item discrimination, sample size ratio, and group ability difference were all significantly related to Type I Error and power for DIF detection. However, these effects were small, and resulted neither in unacceptably high Type I Error nor in unacceptably low power. There were three exceptions to this general finding: 1) the performance of LDFA in detecting nonuniform DIF degraded completely when studied item discrimination was moderate or high, 2) Type I Error rates for classifying DIF were moderately inflated for OLR-uniform when the sample size ratio was 4:1 and 3) Type I Error rates for classifying DIF were moderately inflated for LDFA-uniform when item discrimination was low and there were group ability differences.

Group Sample Size Ratio

Group sample size ratio affected the power of the Mantel, GMH, LDFA, and OLR procedures to detect uniform and nonuniform DIF. For uniform DIF, power was lower when the sample size ratio was 4:1 than when it was 1:1. These effects were small; the power of the Mantel, GMH, and OLR procedures was decreased by 0.025, 0.02, and 0.045, respectively when the sample size ratio was 4:1. The effect of sample size ratio on nonuniform DIF followed a similar pattern, although interactions made the effects more complicated. In general, the higher sample size ratio of 4:1 resulted in lower power for nonuniform DIF for all four procedures. Again, effects were small (around 0.03). For OLR and GMH, the interaction of sample size ratio and item discrimination resulted in some reversals of the trend towards lower power with high sample size ratio.

The effect of sample size ratio on DIF detection has not been well studied. In the only other study which tested the effect of sample size ratio, Tian (1999) found that a 2:1 ratio resulted in a small decrease in the power of some procedures to detect uniform DIF. As previously discussed, large differences in reference and focal group sample size mean that, at each level of total score, there are fewer examinees in the focal group to compare to the reference group; this lowers power.

Sample size ratio did not affect the Type I Error of any of the procedures in cases when there was no DIF. However, the Type I Error of OLR-uniform for differentiating between uniform and nonuniform DIF was somewhat inflated when the sample size ratio was 4:1. Because this is the first time that classification of uniform and nonuniform DIF has been studied, the effect of

smaller differences in sample size is not yet known. The 4:1 ratio used in this study is large, but is realistic for certain testing conditions (e.g., comparing French and English examinees). It is encouraging that the use of such a large ratio resulted in negligible effects in most cases. This suggests that the DIF detection procedures used in this study can safely be used with 4:1 ratios.

Studied Item Discrimination

Item discrimination had a slight effect on the Type I Error and power of most of the procedures, but a large impact on the power of LDFA to detect nonuniform DIF. For Type I Error, there was an interaction between item discrimination and group ability distribution so that Type I Error for the Mantel, GMH, and LDFA was somewhat higher when there was a difference in ability distribution between the reference and focal groups and item discrimination was high. It is important to note that the levels of group ability difference and item discrimination used in the present simulation did not have a serious impact on Type I Error rates. In this study, item discrimination had a small but positive relationship to the power of the Mantel, GMH, and OLR for uniform DIF; as item discrimination increased, so did power. These increases were relatively small, but fairly consistent. This finding is consistent with the fact that the magnitude of uniform DIF is directly dependent on item discrimination so that it increases as item discrimination increases (Chang et al.,1996; Tian, 1999).

The relationship between power and item discrimination was more complicated for nonuniform DIF. For the LDFA, moderate and high item discrimination resulted in extremely low power to detect DIF regardless of whether effect size was included in the decision rules or not. The LDFA should not be used when item discrimination is moderate or high.

For the OLR and GMH with effect size, power was excellent with low discrimination, decreased slightly with moderate discrimination, and was highest at high discrimination. When effect size was not used, power of the OLR and GMH was uniformly perfect and unrelated to item discrimination. Thus, it seems that the actual magnitude of DIF in the simulated data was somewhat smaller for moderate item discrimination. This is due to the fact that the magnitude of nonuniform DIF is dependent on magnitude of item discrimination in the focal group as well as on the difference in a-parameters between the focal and reference groups. Perhaps the predesignated difference in item discrimination was slightly too low for the moderate discrimination condition.

The fact that the level of these effects was relatively small and that Type I Error rates remained low means that OLR and GMH may be used to detect nonuniform DIF, regardless of the level of item discrimination. As mentioned previously, it is actually quite encouraging that these techniques could detect DIF at such very low levels.

More study on the effect of item discrimination on DIF detection for both uniform and nonuniform DIF is needed. For example, a study similar to the one done by Chang et al.. (1996) in which eleven different levels of item discrimination were evaluated would be useful.

Differences in Ability Distribution

The moderate differences in ability between the reference and focal group simulated in this study had a very slight overall impact on DIF detection. For Type I Error, the interaction between group ability difference and item discrimination resulted in significantly higher Type I Error rates for the Mantel, GMH, and LDFA; the magnitude of this effect was very small (around 0.02). This is the first experimental evaluation of the OLR; its Type I Error rates were unaffected by any

conditions. It is possible that it may be more robust than other procedures when group ability differences are larger, but this has yet to be tested.

Group ability differences had very little impact on power; no main effects of group ability were found for uniform DIF. For the GMH and OLR only, there was an interaction between group ability difference and skewness so that if both were present, the power of the DIF detection procedure increased.

Skewness of Ability Distributions

The moderate levels of skewness in group ability simulated for this study had very little impact on the performance of the Mantel, GMH, LDFA, and OLR. Skewness affected neither Type I Error rates nor power to detect uniform DIF. However, skewness did result in slight decreases in the power of the GMH and OLR (with effect size) to detect nonuniform DIF. The interaction between group ability difference and skewness had the opposite effect; power was slightly higher under this condition. These findings are largely consistent with those of Monaco (1997) who reported that moderate skewness had very little impact on DIF detection. However, while she reported no effect, there was a significant, but small effect in this study. This could be due to the fact that she tested different DIF detection procedures.

The Impact of Effect Size

Null hypothesis testing has dominated the behavioural sciences for the last seventy years. Kirk (1996) argued that this longstanding overemphasis on null hypothesis testing has detracted from the main purpose of science which is to interpret outcomes of research, theory, and testing. The tide does seem to be turning; the American Psychological Association recently appointed a

task force to study the desirability of discontinuing null hypothesis testing, and effect size measures are appearing more regularly in published work. Kirk suggested that as an alternative to null hypothesis testing, measures of practical significance which involve point estimates and confidence intervals around the data should be used. These measures of practical significance have several advantages. First, they are in the same unit as the data, which facilitates interpretation. They also clearly show which effects are important and allow the researcher to ignore trivial effects (Kirk, 1996).

The measurement of practical significance is particularly important in the context of DIF studies where there are several potential consequences of being over-inclusive in addition to the well-known consequences of including biased items in tests. For example, Rosnowski and Reith (1999) have shown that items with small amounts of DIF may actually still be very good measures of the construct of interest and that removal of items with small and moderate amounts of DIF may result in tests with lower predictive validity. In practice, items are not automatically thrown out because they show DIF; rather, they are subjected to detailed review. However, it is possible that items which do not really have DIF could be mistakenly discarded even after review (Roussos & Stout, 1996). Furthermore, inclusion of non-DIF items, or of items with insignificant amounts of DIF in such a review would be a waste of time and money.

The inclusion of a measure of effect size can reduce such Type I Errors. One of the a-priori hypotheses in this study was that the effect size measures used in this study would result in significant reductions in Type I Error. This impact was shown to some degree in the present study, although reductions were not as large as had been anticipated. Reductions in Type I Error were largest for the Mantel, decreases ranged from 0.015 to 0.07; the mean reduction was 0.033. The

inclusion of the effect size measure resulted in very modest, but significant mean reductions in Type I Error for the GMH of 0.0079; reductions in Type I Error for the OLR-overall ranged from 0 to 0.013 (M =0.0056). The small magnitude of these reductions is probably due to the fact that the Type I Error of these procedures was already very low and did not vary much across different conditions. As a result, there was very little room for change. It would be interesting to study the impact of effect size in situations which are likely to result in high Type I Error, such as large differences in ability between the reference and focal group combined with very high studied item discrimination.

Inclusion of effect size in the DIF decision rule did have a more substantial impact on Type I Error of LDFA-uniform and OLR-uniform for classifying uniform and nonuniform DIF. Without effect size, the Type I Error for differentiating uniform and nonuniform DIF averaged 0.096 for the OLR-uniform and 0.097 for the LDFA-uniform. Reductions in this form of Type I Error for LDFA-ranged from 0 to 0.07; the mean was 0.025. Type I Error rates for the OLR-uniform when nonuniform DIF was present were reduced by 0 to 0.088; the mean reduction was 0.031. Again, this is probably due to the fact that the Type I Errors for discriminating between different types of DIF were much larger in the first place.

Inclusion of the effect size measure resulted in slight reductions in power as well as in Type I Error. Effect size had little impact on power for uniform DIF. However, it resulted in small reductions in power to detect nonuniform DIF for the GMH and OLR; the mean reduction was 0.066. As mentioned above, this reflects the fact that DIF magnitude in these items was very low (below 0.03 on a 3-point item), although it is difficult to explain why DIF magnitude was lowest in the moderate ranges of studied item discrimination.

In practical terms, the cut-off of 0.03 used in this study means that if 5 items on the total test had DIF at this level, the total difference in test score would only be 0.15 points out of 85. This is relatively trivial, and the decision to ignore such small amounts of DIF is even more conservative than practice at Educational Testing service where items are only sent for review if there is more than a 10% difference in item score. In some situations, however, the consequences of missing a small amount of DIF may be unacceptable. One advantage of using an effect size measure such as the one developed for the present study is that it provides practitioners with all of the evidence that they need to make informed choices on whether the amount of DIF in an item exceeds acceptable limits.

CHAPTER VI
CONCLUSIONS AND RECOMMENDATIONS

This chapter is divided into three sections. In the first section, a summary of study findings and conclusions is presented. In the second section, ideas and recommendations for future empirical research are given. Finally, recommendations for the use of the four DIF detection procedures in applied settings are given and more general conclusions about the use of DIF detection procedures and fairness in testing are presented.

Summary

The Mantel

Prior to the study, it was hypothesized that the Mantel would have good Type I Error control and excellent power for uniform DIF, but that it would be unable to detect nonuniform DIF and could not discriminate between uniform and nonuniform DIF. These hypotheses were largely upheld. The Mantel procedure had excellent Type I Error control across all study conditions, particularly when effect size was included in the decision rule. Furthermore, it had very high power for detecting uniform DIF. However, it could not detect nonuniform DIF and, because it only has one test, cannot be used to discriminate between uniform and nonuniform DIF.

The GMH

Prior to the study, it was hypothesized that the GMH would have very low Type I Error rates, and that it could detect both uniform and nonuniform DIF but that it could not be used to discriminate between uniform and nonuniform DIF. These hypotheses were upheld; the GMH had very low Type I Error rates, and its power for uniform DIF was very high, and equivalent to that

of the Mantel. Furthermore, it had excellent power for nonuniform DIF. The GMH performed well under conditions of low, moderate, and high item discrimination, small group ability differences, moderate skewness, and high reference/focal group sample size ratio.

The LDFA

It had been hypothesized that the LDFA would have good Type I Error control, high power for uniform and nonuniform DIF, and that it could discriminate between uniform and nonuniform DIF. As expected, Type I Error control was good, although the Type I Error rates were somewhat higher for the LDFA than others. It also had excellent power for uniform DIF and was able to discriminate between types of DIF. One other advantage of the LDFA is that it is easier to use than the OLR in cases where the assumption of equal slopes is violated. However, its power for nonuniform DIF was problematic when studied item discrimination was moderate or high, and when there are group ability differences. Power was also decreased when the reference group had four times more examinees than the focal group.

Ordinal Logistic Regression

It was hypothesized that OLR would have excellent Type I Error control, and that it would have excellent power for uniform and nonuniform DIF; it would also be able to discriminate between uniform and nonuniform DIF. These hypotheses were largely upheld. Overall, Ordinal Logistic Regression is the most promising technique evaluated in this study. It had excellent Type I Error control, and high power for both uniform and nonuniform DIF, and can discriminate

between different types of DIF. Furthermore, it was quite robust under the variety of conditions; most factors tested in this study had very little effect on its performance.

One difficulty with OLR is that, if the assumption of equal slopes is not upheld, the user must test each cumulative logit separately. In fact, French and Miller (1996) had previously rejected polytomous logistic regression because of the necessity of testing separate logits. Their main objection was that this made decision making difficult if one logit was significant and the other was not. However, the use of the Bonferoni correction and the effect size measure (which sums across all logits overall measure) overcomes this problem of interpretation.

Effect Size

The effect size measures developed by Aylesworth and Kristjansson (2000) are based on regression predicted group difference in item scores due to DIF. Thus, the unit of measure is intuitively reasonable and comprehensible. Furthermore, because there are separate measures for uniform and nonuniform DIF, they can be used as adjuncts to statistical testing for the identification and classification of uniform and nonuniform DIF. Calculation of these effect size measures is easily done as part of the logistic regression procedure in both SAS and SPSS.

It was hypothesized that inclusion of the effect size measure would reduce the Type I Error of all procedures, and that it would also result in slight reductions in power. Use of these effect size measures did result in decreased Type I Error, although the magnitude of these reductions was not great. They were particularly valuable in reducing the misclassification of uniform and nonuniform DIF. On the other hand, some reduction in power was also found; this was due to the fact that the DIF magnitude was sometimes below the very low threshold in certain conditions.

Thus, the DIF that was missed was that which had relatively little impact on group difference in item score.

Recommendations for Future Empirical Research

1. OLR seems to be the most promising procedure evaluated in the present study. However, this is the first time that it has been studied empirically. Its validity should be further examined under a variety of different study conditions such as small sample size, varying amounts of DIF, and wider levels of studied item discrimination. In particular, it should be compared to GMH again, and also to IRT-based methods. If OLR is comparable to IRT methods, it would be greatly preferred because it is much more practical.

2. The GMH worked as well as the OLR for detecting DIF, at least for large sample sizes. They should be compared in smaller samples, because results of another study (Spray & Miller, 1993) indicate that the power of the GMH may be problematic with small sample sizes. Because the GMH cannot be used to discriminate between uniform and nonuniform DIF, it might be enhanced by adding a test of uniform DIF (e.g., the Mantel) as a second step.

3. Although the LDFA did not perform as well as the OLR or GMH, it is easier to use than the OLR. Furthermore, it has excellent power for uniform DIF, and it can detect nonuniform DIF well under some conditions. Therefore, it should be studied further in order to replicate results thus far, and to determine under which conditions it will and will not work. Perhaps there are more enhancements that can be made to improve its power for nonuniform DIF.

4. It is clear from this study and from other empirical work that the Mantel cannot detect nonuniform DIF. However, nonuniform DIF does sometimes occur in real data (see Table 1).

Therefore, the Mantel should not be studied extensively again unless it is used as a reference to see that the data are performing as expected. It might also be tested for use in a combined procedure with the GMH as the first step, and the Mantel used in the second stage to discriminate between uniform and nonuniform DIF.

5. Thus far, most empirical studies of DIF detection procedures have been limited to one DIF item. In real data, DIF could occur in several items. Thus, the performance of DIF detection procedures should be evaluated when the test contains more than one DIF item. This should be systematically varied by including one, two, three, and more DIF items in the matching test.

6. It proved to be impossible to put large amounts of nonuniform DIF into studied items with this data generation program. It is possible that this represents reality, and that nonuniform DIF just doesn't result in large effect sizes. This issue should be studied further.

7. The type of DIF generated in this particular study was limited to situations with only one crossover. In reality, nonuniform DIF may have several cross-overs. This should be studied further.

Recommendations for Applied Settings

Much of the psychometric literature has been devoted to the development and empirical testing of DIF detection procedures. The guiding principle behind this work is to ensure fair and valid testing for all examinees in real testing situations. Yet, results of the empirical work on DIF detection have been rarely transferred to applied settings; testing for DIF is not yet routinely and systematically done as part of test development and validation. There are several possible reasons for this gap between empirical research and practice. Two of the most likely are: 1) that empirical

research on DIF detection is not reported in a way that makes it easily comprehensible and accessible to practitioners, and 2) much of the literature on DIF detection is devoted to SIBTEST and to IRT methods. These programs are not only expensive to buy, but require extensive data manipulation in order to run. It may be wise to test other more practical methods and to ensure that measurement practitioners are aware that other more easily used (though somewhat less theoretically desirable) methods are available, and that they work well in DIF detection. The present study was designed to test some of these more accessible procedures in order that results from this study might be readily transferred to applied settings. Several recommendations for applied settings have emerged from this study:

1. Graphs should always be used as an adjunct to statistical analyses. Graphical displays lead to a much clearer understanding of the data and may be particularly useful in the classification of uniform and nonuniform DIF, and in understanding how nonuniform DIF affects examinees of different abilities.

2. Before using a DIF assessment procedure, you must check whether the data is appropriate for that procedure. For example, observed score is a good substitute for total test score as a matching variable when the test is reliable, and the data fit a partial credit or General Partial Credit Model. If the test has fewer than 20 items, it might be more appropriate to use a latent trait method rather than an observed score method. Similarly, if items are multiple choice, and guessing is a possibility, then an IRT model might be more appropriate.

3. Dimensionality should be checked before beginning a DIF assessment. If the total score is multidimensional, then it should be divided into unidimensional sub-scales.

4. Always check the assumptions of the statistical procedure you are using. If we hadn't checked assumptions, we would not have known that the assumption of equal slopes was violated, and the OLR would have produced invalid results.

5. In terms of ease of use, all techniques are relatively easy to use, and can be run in SAS or SPSS; specialized computer programs do not need to be purchased. The two modelling procedures are somewhat more time consuming because models require several steps, and, if the assumption of equal slopes is not met, each logit must be tested separately. However, model based procedures allow more flexibility in changing the model to fit the data.

6. Overall, the OLR is the most promising technique tested in this study. It can safely be applied in large samples, even when there are small to moderate group ability differences, and even when item discrimination is high. It also seems to be insensitive to skewed data and to differences in group sample size ratio. Furthermore, the modelling procedure and the effect size calculation used for the present study can be programmed in SAS or SPSS. Thus, it would be recommended as the first choice for DIF detection in applied settings. Those who use it should always test for the assumption of equal slopes, and be prepared to test each logit separately if the assumption is not met. If logits are tested separately, the Bonferoni correction to the chi-square statistic and the overall effect size measure may be used to provide a good indication of the overall level and significance of DIF in the item.

7. The Mantel cannot detect nonuniform DIF. Therefore, it is not recommended for use on its own as a DIF detection procedure.

8. The GMH has excellent Type I error control, and can detect uniform and nonuniform DIF. Thus, it is recommended as a DIF detection procedure. Its one drawback is that it cannot

discriminate between uniform and nonuniform DIF. If empirical research shows that the two-step GMH and Mantel works well, this might be easily transferred to the applied setting.

9. The LDFA is easy to use, but its power for nonuniform DIF is problematic under several circumstances. Therefore, the LDFA should only be used when there are either no, or very small, group ability differences, when the test has fairly uniform item discrimination values, and when the reference/focal group ratio is less than 4:1. Its use in other situations is not recommended.

10 Use of an effect size measure is always recommended; effect size measures show what is really in the data. It makes good sense to routinely obtain an estimate of the predicted difference in score between groups in addition to the statistical testing, and to consider this in decision making. This should ensure that interesting effects are not hidden by small sample sizes and that trivial effects are not deemed statistically significant. The effect size measures developed by Aylesworth and Kristjansson (2000) seem to work well and they can easily be obtained as output in PROC Logistic in SAS. Those who work in an applied setting should know how much DIF is acceptable in that setting, and make judgements about DIF in test items accordingly.

11. In this empirical study, it was necessary to use strict cut-offs for statistical significance and for effect size. However, the rules used to identify items with potential DIF for further study can be much more flexible in an applied situation than those used here in empirical testing; the researcher does not have to adhere to a rigid cut-off.

12. Finally, it is important to undertake DIF detection as part of test validation, particularly now that standardized testing is increasingly being used to select individuals into educational programs and settings, and job training as well to measure the quality of our schools and to compare school boards, provinces, and countries. Item bias may be a significant challenge, but happily, there are

several effective techniques for detecting it. The measures which were evaluated in this study are all included as routines in popular statistical packages. Furthermore, they require little data manipulation, and the statistical code is relatively easy to write. Routine use of such procedures should ensure fair and valid testing for all examinees.

BIBLIOGRAPHY

Ankenmann, R.D., Witt, E.A., Dunbar, S.B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36,* 277-300.

Aylesworth, R., & Kristjansson, B. (2000). Two new measures of effect size for DIF detection with Ordinal Logistic Regression in polytomous items. Unpublished working paper.

Björner, J.B., Kreiner, S., Ware, J.E., Damsgaard, M.T., Bech, P. (1998). Differential item functioning of the Danish translation of the SF-36. *Journal of Clinical Epidemiology, 51,* 1189-1202.

Budgell, G.R., Raju, N.S., Quartetti, D.A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19,* 309-321.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: do item bias procedures obscure test fairness issues? In P. Holland & H. Wainer (Eds.), *Differential item functioning*. (pp. 397-413). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

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

Checkoway, H., Pearce, N., & Crawford-Brown, D. J. (1989). *Research methods in occupational epidemiology. (Monographs in epidemiology and biostatistics; v.13).* New York: Oxford University Press.

Cohen, A.S., Kim, S.H., & Baker, F.B. (1993) Detection of differential item functioning in the graded response model. *Applied Psychological Measurement 17,* 335-35 0.

Cohen, A.S., Kim, S.H., & Wollack, J.A. (1996) An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement 20,* 15-26.

Cohen, J. (1977). *Statistical power analysis for the behavioural sciences.* New York: Academic Press.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data.* Iowa City, Iowa: American College Testing, 97-104.

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

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. & Schmitt, A. P. (1991). *Constructed response and differential item functioning: a pragmatic approach.* (ETS research report 91-47). Princeton, N.J: Educational Testing Service.

Dorans, N.J., Schmitt, A.P, & Bleistein, C.A. (1992). The standardization approach to assessing comprehensive item functioning. *Journal of Educational Measurement, 29*, 309-319.

Ellis, B.B. (1989). Differential item functioning: implications for test translations. *Journal of Applied Psychology, 74*, 912-921.

Fleiss. J.L. (1981). *Statistical methods for rates and proportions (2^{nd} edition).* New York: John Wiley and sons.

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., Clauser, B. E., Mazor, K., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*, 1-18.

Harmon, M, Smith, T.A., Martin, MG., et al. (1997). *Performance assessment in IEA's 3^{rd} International Mathematics and Science Study (TIMSS).* Boston, Mass: International Association for the Evaluation of Student Educational Achievement.

Harwell, M. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 2*, 101-125.

Holland, P., & Wainer, H. (1993). Preface. In P. Holland & H. Wainer (Eds.), *Differential item functioning*. (pp. xii-xv). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Kim, S.H. & Cohen, A.S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22,* 345-355.

Kirk R.E. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement, 56,* 746-759.

Kuehnel, S.M.. (1999). OLOGIT Macro: Logistic Regression with an Ordinal Dependent variable.

Li, H., and Stout, W. (1996). A new procedure for the detection of Crossing DIF. *Psychometrika, 61,* 647-677.

Lord, F.M. (1955). A summary of observed test score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement, 15,* 383-389.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum.

Longford, N.T., Holland, P.W., Thayer, D.T. (1993). Stability of the MH D-DIF statistics across populations. In P. Holland & H. Wainer (Eds.), *Differential item functioning.* (pp. 397-413). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Marshall, S.C., Mungas, D., Weldon, M. Reed, B. & Haan, M. (1997). Differential item functioning in English- and Spanish-speaking older adults. *Psychology and Aging,* 718-725.

Mantel, N. (1963). Chi-square tests with one-degree of freedom: extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58,* 690-70 0.

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.

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.

Monaco, M.K. A Monte Carlo assessment of skewed theta distributions on differential item functioning indices. *Dissertation Abstracts International: Section B: the Sciences and Engineering.* Vol 58 (5-B), Nov 1997, 2746.

Nandakumar, R., & Yu, F. (1996). Empirical validation of DIMTEST on nonnormal ability distributions. *Journal of Educational Measurement, 33,* 355-368.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: a framework for classification and evaluation. *Applied Psychological Measurement, 19,* 23-27.

Price, L.R., Oshima, T.C. (1998). Differential Item Functioning and Language Translation: A cross-national study with a test developed for certification. Paper presented at the 1998 Annual Meeting of the American Educational Research Association, San Diego, CA.

Raju, N.S., van der Linden, W.J., and Fleer, P.F. (1995). An IRT-based internal measure of test bias with application for differential item functioning. *Applied Psychological Measurement, 19*, 353-368.

Roznowski, M. & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: do biased items result in poor measurement? *Educational and Psychological Measurement, 59*, 248-269.

SAS Institute (1995). *Logistic Regression Examples Using the SAS® System, Version 6, First Edition.* Cary, NC: SAS Institute, Inc.

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

Simon, M. (1994). Differential item functioning: applicability in a bilingual context. In D. Laveault, B. Zumbo, M. Gessaroli, & M. Boss (Eds.), *Modern theories of measurement: problems and issues*. (pp. 163-180). Ottawa, Ontario: University of Ottawa.

Smith, L.L & Reise, S.P. (1998). Gender differences on negative affectivity: an IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale. *Journal of Personality and Social Psychology, 75*, 1350-1362.

Spray, J. (1992). *Psych.For.*

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

Stokes, M.E., Davis, C.S., & Koch, G.G. (1995). *Categorical data analysis using the SAS system.* Cary, N.C: SAS Institute, Inc.

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

Streiner, D.L. & Norman, G.R (1995). *Health measurement scales. A Practical Guide to their Development and Use.* Oxford: Oxford University Press.

Tanzer, N. K. (1995). Cross-cultural bias in Likert-type inventories: perfect matching factor structures and still biased. *European Journal of Psychological Assessment, 11*, 194-201.

Thissen, D., Steinberg, L, & Wainer, Y. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer, & H. Braun (Eds.), *Test Validity* (pp147-170). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L, & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. Im P.W. Holland & H. Wainer (Eds.), *Differential item functioning: theory and practice* (pp67-113).

Tian, F. (1999). *Detecting differential item functioning in polytomous items.* Unpublished doctoral dissertation, Faculty of Education, University of Ottawa;

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

Welch, C.J. & Miller, T.R. (1995). Assessing differential item functioning in direct writing assessments: problems and an example. *Journal of Educational Measurement, 32,* 163-178.

Woodward, J.L, Auchus, A.P., Godsall, R.E, and Green, R.C. (1998). An analysis of test bias and differential item functioning due to race on the Mattis Dementia Rating Scale. *Journal of Gerontology: Psychological Sciences, 53B,* P370-P374.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning.* (pp. 397-413). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Zumbo, B. D. *A handbook on the theory and methods of differential item functioning (DIF). Logistic regression modeling as a unitary framework for binary and Likert-type item scores.* (1999). Ottawa, Ontario: Directorate of Human Resources Research and Evaluation, Department of National Defence.

Zwick, R., Donoghue, J., & 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 Behavioural 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.

Appendix One
The Containers Item from the
Third International Science and
Mathematics Study

From Harmon, Smith, Martin et al, 1997

CONTAINERS

In the Containers task, students were given three containers of different insulating capacity, for example, metal, ceramic, and plastic, and were asked to find out which one would keep a hot drink warm for the longest time. They also received thermometers, a clock, a piece of card to use as a fan, and a supply of hot water. The students were instructed to pour a measure of hot water into each of the containers, and to take the temperature in each one over a ten-minute interval. They were provided with a pre-designed data table in which to record their observations. This task assessed students' ability to make and record measurements of temperature and probed their understanding of the concept of insulation. Figure 1.9 presents the task with sample student responses and scoring criteria for a fully-correct response. This task was administered to fourth-grade students only.

In general, this was a difficult task for fourth graders. Although most students in most countries were able to use a laboratory thermometer, in many cases, the data gathered were incomplete or contained small inaccuracies in measurement (Table 1.10, Item 1 – average percentage scores: 91% and 56%). Students did reasonably well in identifying the container that kept water hottest (Item 2 – average percentage score: 48%), but almost none could explain insulating capacity in terms of the materials from which the containers were made.

An interesting misconception appeared when students were asked to apply their findings to a different situation – that of keeping ice cream cold. While 15% of students internationally (Item 4) recognized that the container that was best for keeping a hot drink warm would also be best for keeping ice cream cold, almost none could explain why (Item 5). About one-quarter of the students seemed to see the ice cream as an opposite case, explaining that the container in which the temperature of a hot drink declined most rapidly would be the one to keep ice cream cold the longest.

FULL-TASK EXAMPLE AND SCORING CRITERIA – FOURTH GRADE

INTRODUCTION TO TASK

CONTAINERS

At this station you should have:

Three containers (or cups) marked A, B, C
Three thermometers
A clock or watch
A container with very hot water. BE CAREFUL. NOT TO SPILL HOT WATER.
Pieces of card to use as a fan if you wish
A roll of paper to wipe up spills
A measuring cup

Read ALL directions carefully.

Your task:

Find out which of the containers will keep a hot drink warm for the longest time.

This is what you should do:

- Place a thermometer in each of the containers BEFORE the hot water is poured in. Your teacher will pour the hot water when you are ready. BE CAREFUL. THE WATER IS VERY HOT.

- Measure the temperature on each thermometer as soon as the hot water is poured in.
- Write these measurements and the time in the table on the opposite page.
- Now you will take measurements over a total of 10 minutes.
  - → Decide how often to read each thermometer.
  - → Write your measurements in the table on the opposite page.

page 1 TASK S6-P1

ITEMS 1, 2, AND 3

1. Table of Measurements:

Time	Temperature of Container A	Temperature of Container B	Temperature of Container C
2:28	0	0	0
2:32	50	60	50
2:35	50	60	50
2:36	50	60	50
2:38	50	60	40
2:39	50	50 ½	40

2. Look at the table. Which container keeps a hot drink warm for the longest time? cupB plastic

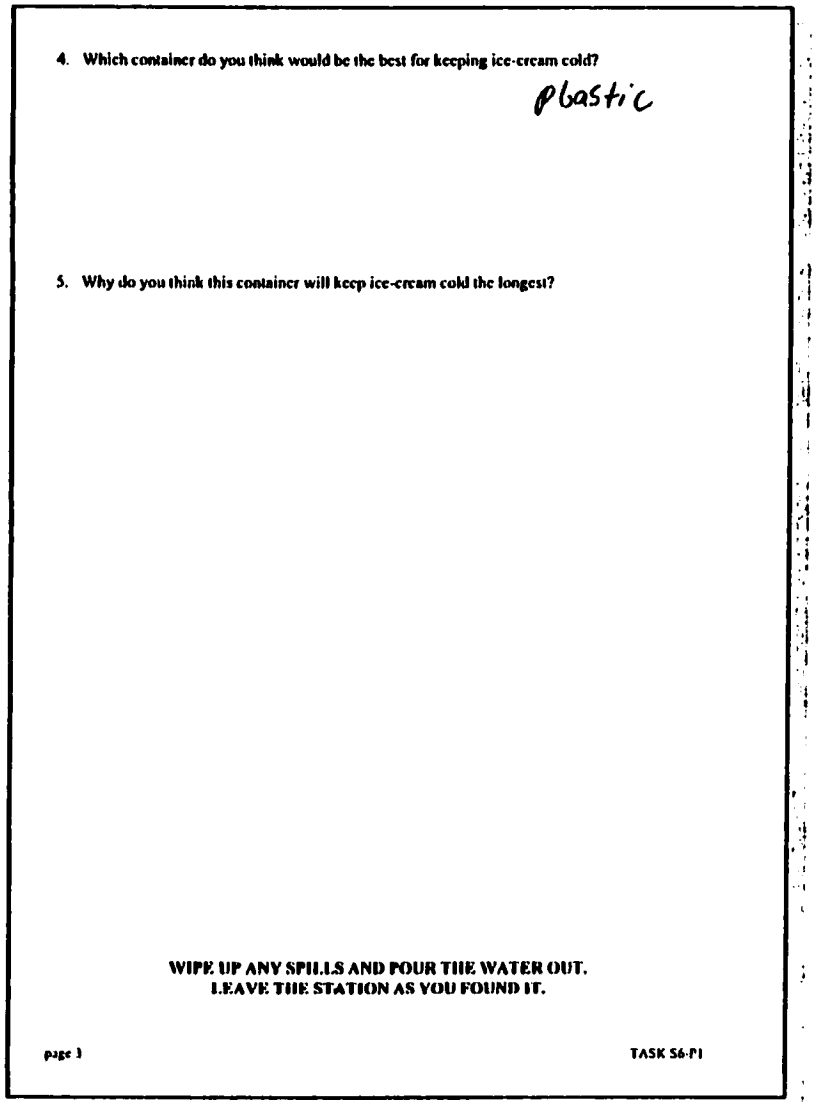
3. Why do you think this container was best for keeping a hot drink warm?

Because it holds the heat

Please turn the page.

TASK S6-P1 page 2

ITEMS 4 AND 5



4. Which container do you think would be the best for keeping ice-cream cold?

plastic

5. Why do you think this container will keep ice-cream cold the longest?

WIPE UP ANY SPILLS AND POUR THE WATER OUT.
LEAVE THE STATION AS YOU FOUND IT.

page 3 TASK S6-P1

CRITERIA FOR FULLY-CORRECT RESPONSE

Item 1 - Measure temperatures and record data in table. Student is scored both on proper use of the thermometer and on the quality of data gathering.

Ability to use thermometer. Does not require assistance in proper use of the thermometer (Based on adminstrator notes on any special assitance provided.)
Total Possible Points: 1

Quality of data gathering. i) Records times and temperatures for 5 or more temperature points per container. ii) Times cover full 10-minute range. iii) Trend in the temperature is reasonable: temperature declines with time in one or more of the cups. (One cup may be too well insulated to give measurable declines in 10 minutes.)
Total Possible Points: 3

Item 2 - Identify container that keeps hot drink warm longest. i) Identifies correct container (based on administrator notes). ii) Container identified is consistent with the data in table.
Total Possible Points: 2

Item 3 - Explain why container retains heat. i) Relates material of containers to their ability to retain or transfer heat. ii) Includes comparison of different containers based on heat transfer. iii) Logically applies any additional relevant information (stirring, thickness of container, size differences, etc.).
Total Possible Points: 2

Item 4 - Predict best container for keeping ice cream cold. Identifies the same container that best keeps hot drink warm.
Total Possible Points: 1

Item 5 - Explain why container keeps ice cream cold. i) Relates material of containers to their ability to retain or transfer heat. ii) Includes comparison of different containers based on heat transfer. iii) Logically applies any additional relevant information provided (stirring, thickness of container, size difference, etc.).
Total Possible Points: 2