Distribution and Power of Selected Item Bias Indices: A Monte Carlo Study

Abdul K. Ibrahim

Thesis submitted to
the School of Graduate Studies and Research
in partial fulfillment of the requirements for the Ph.D.
degree in Education

University of Ottawa

1992

ISBN 0-315-85766-8

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

Abstract

This study examines the following DIF procedures - Transformed Item Difficulty (TID), Full Chi-Square, Mantel-Haenszel chi-square, Mantel-Haenszel delta, Logistic Regression, SOS2, SOS4, and Lord's chi-square under three sample sizes, two test lengths, four cases of item discrimination arrangement, and three item difficulty levels. The study is in two parts: The first part examines the distributions of the indices under null (no bias) conditions. The second part deals with the power of the procedures to detect known bias in simulated test data. Agreements among procedures are also addressed. Lord's chi-square certainly appears to perform very well. Its detection rates were very good, and its percentiles were not affected by discrimination level or test length. In retrospect, one would like to know how well it might do at smaller sample sizes. When the tabled values were used, it performed equally well in detecting bias and improved in reducing false positive rates.

Of the other indices, the Mantel-Haenszel and the logistic regression indices seemed the best. Camilli chi-square had a number of problems. Its tabled values were not at all useful for detection of bias. The TID was somewhat better but does not have a significance test associated with it. One would need to rely on baseline studies, if one were to use it.

For uniform bias either Mantel-Haenszel chi-square or logistic regression would be recommended, while for nonuniform bias logistic regression would be appropriate. It is interesting to note that Lord's chi-square was effective for detecting either kinds of bias.

We have been told that sample size is related to chi-square values. For each of the chi square indices the observed values were considerably lower than tabled values. Of course, these were conditions where no bias was present except that which might be randomly induced in data generation. Perhaps it is those instances where bias is truly present that larger sample sizes allow us to more easily identify biased items. Certainly the proportions of biased items detected was greater for large sample sizes for Camilli chi-square, Mantel-Haenszel chi-square, and logistic regression chi-squares.

Acknowledgements

I wish to start by expressing my sincere gratitude to my supervisor, Dr. Marvin Boss, whose advice, guidance, suggestions and criticisms were of extreme value in helping me through this piece of work. I always relied on his insightful guidance whenever I seemed to have been trapped in a dead end.

To Dr. M. Gessaroli I wish to express my sincere thanks for the helping hand you extended to me whenever I needed one.

I owe a lot of gratitude to Michael Brabant and Len Fleming for the help they gave me with writing up and/or compiling some of the computer programs I needed for the study. I really and truly appreciate the help you gave to me in my hours of need.

The typing of this paper was done by Ronalda Rino. There are no words with which I can express my thanks to you, Ronalda, for the amount of work you had to do in a moment's notice. I will always remember your effort in seeing this work completed.

Last, but not least, I wish to say a big 'thank you' to my wife Anna and my two sons Abdul Jr. and Ahmed. I do appreciate the love, support and understanding you gave to me. I pray that I will be there for you when your turn comes for such love and support.

Thanks to all the many friends I have made since I came to the University of Ottawa.

Table of Contents

List of Tables

List of Tables (Continued)

Page

List of Tables (Continued)

Page

CHAPTER I

Introduction

One of the most recent developments in the world of testing is an explosion of the concern for unbiased tests. Test developers and test users are presently very sensitive to the way their tests perform in different subgroups found in the test's target population. This sensitivity manifests itself in the effort put into action to ensure that the test behaves in a similar way in each of the subgroups of the target population. The exercise takes the form of either examining the test as a whole or examining individual items in the test with respect to the way they function in the different subgroups. When the problem is addressed from the whole test angle, this is referred to as a "Test Unfairness Study" or a "Test Bias Study". When the problem is addressed from the individual test item angle, it is referred to as an "Item Bias Study". This study concerns the latter.

To facilitate item bias studies, researchers like Swaminathan and Rogers (1989); Holland and Thayer (1986); Linn et al., (1981); Shepard, Camilli and Averill (1981); Lord (1980); Rudner (1977); Angoff and Ford (1973) (to name a few) have developed certain indices which can be used to determine whether an item is behaving differentially in different subgroups in a target population. Such indices are broadly labelled "statistical procedures for detecting bias in test items". Another way in which item bias studies are done is through the use of experts in the knowledge area addressed by the test. These experts examine the test item's format and

content and make a judgement as to whether the items is biased against any of the subgroups in the target population. This method is labelled the "Judgemental Method". This study does not examine this non-statistical approach.

Over the years, there has been a proliferation of statistical item bias detection procedures. Each of these procedures takes the form of a mathematical model which uses some parameters from the item and the test to produce an index whose value is used to determine whether the item is biased. The unfortunate thing about these procedures, is that their bias detection results do not always agree. This phenomenon has given rise to series of studies which try to compare some statistical procedures. Based on results from such studies, strengths and weaknesses of different procedures have been gradually surfacing.

The current study is in part a comparative study, in terms of examining the power of detecting biased items for different statistical procedures. The study will also attempt to examine the distributions of the indices of the different procedures when there is no bias, other than chance bias, present in the test data.

The distribution study will examine the effect of certain variables of interest (test length, sample size, and item discrimination) on the indices produced by the different procedures. The procedures in this study include three procedures based on item response theory, IRT, (these are Lord's chi-square, SOS2 and SOS4) and five non-IRT procedures - the logistic regression (LR) procedure, the Mantel-Haenszel (MH) chi-square and

delta procedures, the Full chi-square procedure and the Transformed Item Difficulty (TID) procedure.

The review of literature is presented in Chapter II. Chapter III gives a description of the study. The Methodology used in the study is presented in Chapter IV. The results are presented and discussed in Chapter V. The conclusions of the study are presented in Chapter VI.

CHAPTER II

Review of Research

Tests are normally designed to aid in some decision making processes. In most cases, members of the target population, for whom a test is designed, vary on certain characteristics such as gender and social or cultural backgrounds. When such subgroups exist within the target population of a test, the whole test or the individual items in the test may function differentially for each identifiable subgroup. This concept of differential functioning of a whole test, or the items in the test, may be as old as the introduction of testing itself. However, the issue never received the attention it is due until the late 1960's and early 1970's. Since then, an appreciable proportion of educational research has been directed toward answering questions concerning differentially functioning tests and/or test items. In addition to this proliferation of research, test developers and test users now have the added responsibility of ensuring that a test , destined for a target population that displays gender, social or cultural differences in subgroups, does not give undue advantages for success to members of a particular gender, social or cultural subgroup over members of another.

In contemporary society, it is difficult, if not impossible, to find a target population for a test that does not display at least one of these gender, social or cultural differences in its

subgroups. This implies that for the majority of tests, test developers and test users must address the issue of differential functioning at either the test level or at the individual item level. To help test developers and test users answer some of their questions concerning differential functioning, two types of research works surfaced: those that address the issue of fairness of the testing instrument as a whole and those concerned with the individual items.

The issue of fairness of a test is of major concern when the test is to be used in selection purposes. In such a situation the test is used as a predictor of some future performance (i.e. the criterion). If such a test should predict the criterion well for one subgroup and poorly for another subgroup, within the target population, it will give an undue selection advantage to the former subgroup while it gives an undue disadvantage to qualified members of the latter subgroup. Such a test is said to be a "biased test". This situation is also sometimes referred to as "unfairness". According to Rudner, Getson, and Knight, (1980), researchers who address the issue of test bias attempt to answer the question: "Does the test either unduly favour or impede examinees from different parts of the country or from different background in relation to their performance on the criterion?" (p. 214).

Where the concern for differential functioning is addressed at the individual test item level, rather than at the whole test level, the term "item bias" is generally used in the literature. The definition of "item bias" varies with the research method used

to address it. However, the most commonly accepted definition is the one used by Shepard, Camilli, & Averill (1981): "An item is classified as biased if two examinees of comparable abilities but from different subgroups within the target population do not have the same probability of success on the item." This form of bias (i.e., item bias) is the focus of this study.

Item Bias

The main concern with item bias, in the field of testing, has to do with the identification of biased items in a test and the correction or elimination of such items depending on whether the items can be corrected or not. It can be argued that, for any given test containing more than one item, the characteristics of the test as a whole are not necessarily the same as the sum of the characteristics of the individual items that constitute the test. In any case, for the relationship between test bias and item bias, it can be assumed that a test, made up of unbiased items for all the subgroups in the target population, will itself be an unbiased test for all the subgroups of interest. Lord (1980) suggests that a test would be completely unbiased if each of the items in the test has exactly the same item response function in every subgroup. This implies that if the test is made up of items, all of which measure the same latent trait, as long as the items are unbiased, the test will also be unbiased. In this respect, a study that

addresses bias at the item level will have an advantage over one that addresses bias at the whole test level.

To be able to detect biased items, certain procedures must be defined. Item bias detection procedures found in the literature can be divided into two broad classes: judgemental procedures and statistical procedures. In judgemental procedures, the test item is usually examined by one or more "expert judges". The judges are requested to examine the item's format and content for clues that may render the item more difficult for members of one subgroup than for members of another subgroup within the test's target population (Tittle, 1982). The major problem with this procedure is the subjectivity of the expert judges.

Statistical procedures are based on mathematical modeling that utilize parameter estimates obtained from some form of analyses of the test scores for the subgroups of interest. The mathematical formula for each statistical procedure results in an estimator, the value of which is used to determine bias. Some of the indices obtained from such statistical procedures (e.g. the logistic regression procedure and the Mantel-Haenszel chi-square procedure) have known statistical distributions with definite tests of significance associated with them. A number of others (e.g. the Item Response Theory - Area Between procedure and the Transformed Item Difficulty procedure) have no known statistical tests of significance associated with them. For each procedure in the latter set, it is necessary to define some objective method that will be used to determine whether the value of the index indicates DIF or

not. One such method is the baseline study method (see, for example, Shepard, Camilli, and Williams, 1985). So far, very little, if any, similarity is known to exist between results obtained from the judgemental and the statistical approaches (Burrill, 1982). However, there is agreement in the literature that the results from the two approaches should be used to complement each other.

Item bias cannot be fully defined in statistical terms. It is more of a philosophical and emotional concept, (Burrill, 1982). Statistical procedures cannot be wholly relied upon to detect item bias. The content and format of an item, flagged by a statistical procedure, have to be carefully examined to establish if indeed the problem with the item could be classified as biased. When an item is flagged by such statistical procedures, the only valid statement that could be made about the item is that it is performing differentially in the population subgroups being studied. Some authors have suggested the use of more appropriate terms, "DIFFERENTIAL ITEM PERFORMANCE, (DIP)" (Welch et al., 1987; Wright, 1986) or "DIFFERENTIAL ITEM FUNCTIONING (DIF)" (Holland & Thayer, 1986), in place of the term "ITEM BIAS", when statistical procedures are used. The term Differential Item Functioning or DIF shall be adopted in this study.

Concern for bias-free tests in contemporary society has given rise to a proliferation of statistical procedures for detecting DIF. Unfortunately, the results obtained from the different procedures do not display a strong enough agreement. This lack of

agreement has resulted in an appreciable number of studies each aiming at identifying procedures that are best indicators of DIF (e.g. Rudner, Getson & Knight, 1980; Shepard, Camilli & Williams, 1985).

Statistical DIF detection procedures that have been developed to date can be divided into CONDITIONAL and UNCONDITIONAL procedures. In conditional procedures, the samples of examinees from the subgroups of the target population are matched on ability. In unconditional procedures, the samples are not matched with respect to ability. Unconditional procedures have been criticized for confounding bias with difference in ability between the subgroups. In studies that compare DIF detection results obtained from different procedures, with the exception of the Transformed Item Difficulty (TID) procedure, unconditional procedures have always performed poorly (Merz & Grossen, 1979; Rudner, Getson, and Knight, 1979, 1980).

In DIF studies, situations arise where an item displays DIF in favour of one subgroup at lower ability levels and DIF against the same subgroup at higher ability levels. When this happens, the item is said to display "NONUNIFORM DIF". If an item displays DIF against one subgroup at all ability levels, the DIF is referred to as "UNIFORM DIF". Some statistical DIF procedures (e.g. the Mantel-Haenszel Chi-Square procedure) are known to be sensitive to Uniform DIF only.

Statistical DIF detection procedures differ not only in their conceptualization of DIF but also in their theoretical soundness,

statistical complexity, sample-size requirements, and cost of implementation (Ironson, 1982; Shepard, Camilli, & Averill, 1981). This is still true today. Over the years, attempts have been made to examine the theoretical soundness, efficiency in identifying biased items, and reliability as determined by the degree to which a given procedure can replicate its DIF results. Most of the DIF detection procedures have been found wanting in a few or all of those properties.

DIF detection procedures based on Item Response Theory (IRT) models, especially the three-parameter IRT model, are considered the most theoretically sound procedures (Ironson & Subkoviak, 1979; Rudner & Knight, 1980; Shepard, Camilli & Averill, 1981; Ironson, 1982). This is mainly due to the fact that these procedures do not confound real mean differences in groups' performances with DIF. However, the cost of implementation, statistical sophistication, and large-sample-size required for estimation of item parameters put the three-parameter IRT procedures beyond the reach of many practitioners in the field of education. Most of the three-parameter DIF detection procedures can also be used with the less sophisticated one- and two-parameter IRT models, as long as the response data fit these models. In the literature, procedures based on the one- and three-parameter IRT models have received more prominence than those based on the two-parameter model, the one-parameter model is much more tolerant with respect to small sample sizes. Traub (1983) states that in real testing situations where multiple-choice items are used, some amount of guessing is always

present, and that it is not likely that all the items in a test will have equal discrimination as is assumed by the one-parameter model. On the basis of these points raised by Traub, the one-parameter model can be criticized as not representing reality. On the other hand, the three-parameter model is much more of a representation of reality because it provides for guessing and does not assume uniform discrimination among items.

By their nature and the parameters used in obtaining the values for their indices, some three-parameter IRT DIF detection procedures can be modified for use with the one-parameter IRT model (e.g. procedures based on the sum of squared differences, in probability of success, between the subgroups under study). In the literature, however, two main DIF detection procedures are associated with the one-parameter IRT model. These two procedures are based on:

I) the examination of differences in item difficulty levels for item across the two subgroups of interest (Draba, 1977; Wright, Mead & Draba, 1976).

II) the examination and comparison of the item/model fit statistics obtained for each of the two subgroups of interest (Wright, Mead & Draba, 1976; Durovic, 1975).

The most prominent among the three-parameter IRT DIF detection procedures found in the literature include:

I) Linn, Levine, Hastings, and Wardrop, (1981); Shepard, Camilli & Averill, (1981); Rudner, (1977): they defined indices based on the area vetween the item characteristic curves (ICC's) of

the two subgroups under study. This method utilizes the fact that since the IRT item parameters are not sample dependent, given allowances for chance variation, the ICC's for two equivalent (i.e. equal in ability) groups, for an item, should be identical.

II) Shepard et al., (1984); Linn et al., (1981): they defined indices based on the sum of squared differences between the probabilities of a correct response, to an item, for the two groups under study.

III) Lord (1980) defined an index which is an asymptotic significance test that simultaneously compares the differences of the discrimination (a) and the difficulty (b) parameters between the groups under study.

IV) Linn and Harnisch (1981): they proposed an index based on the average difference between the empirically obtained probability of success and the expected probability of success of the minority group (i.e. the smaller of the two groups under study). This difference measures the degree to which members of the minority (or target) group performed better or worse than expected. They also defined a second index which is the standardized form of the first.

V) Linn et al., (1981): they proposed a procedure which involves the visual examination of each group's ICC, plus and minus twice its standard error, plotted on a common graph.

VI) Hulin et al., (1982, 1983); Bougon and Lissak (1981) proposed indices based on the empirical item response function (EIRF) for the two groups under study.

Over the years, a good number of non-IRT DIF detection procedures have appeared in the testing world. Some of these never reached any level of prominence because of theoretical flaws detected in their designs. Some such procedures are: Scheuneman's (1979) chi-square procedure, a point-biserial correlation procedure (Green & Draper, 1972), a partial correlation procedure (Stricker, 1982) and procedures based on multivariate factor structure analyses (Lei & Skinner, 1983; Rock & Werts, 1979; Merz, 1973, 1976; Green & Draper, 1972).

Non-IRT DIF detection procedures that have featured prominently in the item bias literature include:

I) The Transformed Item Difficulty (TID) procedures (Rudner, Getson, & Knight, 1980; Echternacht, 1974; Angoff & Ford, 1973; Angoff, 1972). The original TID procedure is based on the examination of bivariate plots. Each point in the plot has co-ordinates that are delta scores obtained from the transformation of the "classical" item difficulty (p-value) computed for each of the groups under study.

II) The Full Chi-Square procedure proposed by Camilli (1979). Over several ability groupings, this procedure compares the observed frequencies of examinees answering an item correctly and examinees answering the item incorrectly to their

respective expected frequencies for answering the item correctly or incorrectly in each of the groups under study.

III) Dorans and Kulick (1986) defined two indices for which the total test score is divided into total test score classes. The item difficulty (p-value), for examinees who fall into a particular total test score class, was determined for each of the groups under study. The difference in p-value in each score class was obtained. These differences, over all the test score classes, were standardized and summed to define the first index. The second index is simply a weighted root mean squared transformation of the first.

IV) The Mantel-Haenszel (M-H) chi-square procedure, proposed by Holland and Thayer (1986). This is a test statistic for the Mantel-Haenszel odds ratio.

V) The Mantel-Haenszel (M-H) Delta procedure, proposed by Holland and Thayer (1986). It is a log transformation of the Mantel-Haenszel odds ratio which shows amount and direction of bias.

VI) The loglinear procedures proposed by VanDer Flier, Mellenbergh, Ader and Wijn, (1984); Mellenbergh, (1982). Basically, a loglinear procedure tests the fit of models by means of a maximum likelihood ratio statistic, G, which is asymptotically distributed as chi-square, and is defined in terms of the log of the ratio of the observed cell frequency to the expected cell frequency from a three dimensional (score x group x response) contingency table.

VII) The logistic regression procedure proposed by Swaminathan and Rogers (1989). This procedure models the probability of a correct response to a biased item to depend on both group membership and ability using the standard logistic regression model for predicting a dichotomous dependent variable from given independent variables. With this procedure, an unbiased item will have identical regression curves (i.e. equal intercepts and equal slopes) for the two groups under study.

Previous Studies

Each statistical DIF detection procedure is in the form of a mathematical formula which defines an index, the value of which is used to determine DIF. Because mathematics allows for different approaches to the solution of any given problem, the different mathematical formulae associated with the different statistical DIF procedures can be justified. However, in a normal mathematical problem solving situation, the different procedures are expected to converge on the correct solution to the problem. Unfortunately, this is not the case with all the currently known DIF procedures. They do not always converge on the same solution. The search for convergence of results has been the main feature of studies comparing statistical DIF detection procedures (Hambleton & Rogers, 1989; Merz & Grossen, 1979; Rudner, Getson, & Knight, 1980; Seong & Subkoviak, 1987; Shepard, Camilli, & Averill, 1981; Shepard, Camilli, & Williams, 1984, 1985; Spray, 1989). At the heart of this

search for convergence of results lies the more important reason which has to do with concern for practitioners in the field of education, the ultimate users of these procedures. From the practitioner's point of view, even though efficiency and theoretical soundness of a procedure is important, there are other factors that must be considered. These include cost of implementation, ease of understanding, interpretability to lay public and other practical requirements such as sample size and test length.

Two main types of studies, comparing DIF detection procedures, are found in the literature. These shall be labelled SIMULATED-DATA STUDIES (SDS) and EMPIRICAL-DATA STUDIES (EDS). In SDS, test data are generated in such a way so as to establish, a priori, both the nature and amount of bias present in the test data. DIF detection procedures that are being compared are then applied to the data to determine the effectiveness of each of the procedures in identifying the known bias in the test data. The results obtained from the different procedures are then used to evaluate the relative efficiency of each procedure with respect to identifying the known bias in the test data. The DIF detection similarities among procedures, the amount of false-positive identification (i.e. identifying as biased, an item which was generated unbiased), and the amount of false-negative identification (i.e. identifying as unbiased, an item which was generated biased) could also be determined. Some studies belonging to the SDS group are Merz and Grossen, (1979); Rudner, Getson and Knight, (1980); Shepard, Camilli and Williams (1985); Rogers and Swaminathan, (1989, 1990).

In each of the SDS studies, mentioned above, the true state of bias in the test data was known because the data were generated that way. In Merz and Grossen, (1979), Rudner, Getson and Knight, (1980), and Shepard, Camilli and Williams (1985), DIF detection procedures with no known statistical test of significance associated with their indices were included (e.g. the TID procedure and some IRT procedures). Whenever such DIF procedures are used, some objective way of establishing cut scores (e.g. the baseline study method) for the indices must be employed in the study and reported. Unfortunately, of the studies using such procedures, mentioned above, only the report of Shepard, Camilli and Williams (1985) mentioned the use of such an objective method for establishing cut scores. The studies by Rogers and Swaminathan (1989, 1990) compared only two DIF procedures, the Logistic Regression (LR) procedure and the Mantel-Haenszel (MH) procedure. Both of these procedures have known statistical tests of significance associated with their indices.

In EDS, attempts are made to compare different DIF detection procedures on the basis of their ability to detect bias in test data obtained from the performances of actual people on actual test items. Two sub-classes of EDS are identifiable in the literature. In one sub-class, the researcher attempts to build a "known" amount and nature of bias into the test. This is done mainly by getting experts to construct test items whose contents and/or format they (the experts) believe are biased against members of one of the sub-groups of the target population. This EDS sub-class assumes that

the true nature and amount of bias, in the data obtained from such a test, are known. Like studies in the SDS category, studies in this EDS sub-class proceed to evaluate the different DIF procedures in terms of their accuracy in detecting the "known" bias in the data, their similarities in bias detection results, their false-positive identifications and their false-negative identifications. Examples of studies falling into this first sub-class of EDS are Ironson et al., (1984), Subkoviak et al., (1984), and Seong and Subkoviak (1987).

The second sub-class of EDS comprises of studies in which no attempts are made to build known bias into the test that is being studied. Effectively, the amount and nature of bias in the test data are accepted as unknown. With the exception of a few studies (e.g. Shepard, Camilli & Williams, 1985) in this EDS sub-class, the only meaningful comparisons that could be made are in terms of the agreement and disagreement among the DIF procedures. When the true state of bias in the test data is unknown, even when different DIF procedures flagged the same item, it is difficult to accept such an item as biased with full confidence. An examination of the item's content and format will be necessary to confirm the quality of the item. It is also possible that the most meticulous examination of such flagged item's content and format will give no clue of bias. All the flagging of the item tells us is that the item is behaving differentially in the sub-populations. This differential behaviour may be due to bias or to some other factor (e.g. the misspecification of the latent ability space for the test - see Ackerman,

1991). Some studies in this sub-class are Vaillancourt, (1984), Wright (1986), Rogers and Hambleton (1988), and Schutz et al., (1989).

In some other studies, like Shepard, Camilli and Williams (1985), which also belong to the second EDS sub-class, though the true state of bias in the data is not known a priori, the researcher(s) first use a theoretically sound DIF procedure (e.g. one of the well tested DIF procedures based on the three-parameter IRT model) on the data. The result obtained from the use of this procedure is then accepted as a full definition of the true state of bias in the data. Based on this established true bias state, other DIF procedures are judged in terms of their ability to replicate the established "true" state of bias in the data. With such a "true" bias state established, the efficiency, false-positive identification and false-negative identification of the other DIF procedures could be determined.

Simulated Data Studies

Rogers and Swaminathan (1990) examined the chi-square distributional claims and compared the powers of identifying biased items of both the Mantel-Haenszel (MH) and the Logistic Regression (LR) DIF procedures. Sample size and model-data fit were identified as factors that might affect the distributional claims of the indices for the two DIF procedures. For biased item identification power, the researchers identified six factors that might affect the results from both the MH and LR procedures. The six factors were

model-data fit, sample size, test length, shape of test score distribution, size of bias, and proportion of biased items in test.

The LR statistic is expected to be distributed as a chi-square with 2 degrees of freedom (with mean and standard deviation each equal to 2.0)., The MH statistic is expected to be distributed as a chi-square with 1 degree of freedom (with mean equal to 1.0 and standard deviation equal to $\sqrt{2.0}$). In the distributional study, two sample sizes (250 and 500 subjects per group) were crossed with two model-data fit levels (good-fit and poor-fit). Five selected items with no bias induced, and with varying levels of difficulty and discrimination, were used in each of the four experimental conditions. Data were generated for each of the five items, under each condition, and replicated 100 times. The LR and MH statistics were calculated and empirical sampling distribution constructed separately for each statistic. The Kolmogorov-Smirnov test (in the form of a K statistic) was then performed on each statistic's distribution for each item under each condition.

Three out of the twenty distributions of the LR statistic displayed a significant K at the .05 level while six out of the twenty distributions for the MH statistic had significant K's at the .05 level. The MH statistic also displayed a slight effect due to sample size. For the 250 per group sample data, 3 out of 5 and 2 out of 5 cases, for the good-fit and poor-fit conditions respectively, displayed significant K's, while 0 out of 5 and 1 out of 5 had significant K for the corresponding 500 per group data. The 6 significant K cases for the MH index were different from the

3 significant K cases for the LR index. Also the significant K cases for the MH index did not correspond in the different conditions (e.g. items 1, 4 and 5 in one condition, items 1 and 4 in another condition and item 2 in a third condition displayed significant K's).

In the "power of identification of bias" study 2 levels (good/poor) of model-data fit, 2 levels (250 per group/500 per group) of sample size, 2 levels (40 items/80 items) of test length, 2 levels (normal/skewed) of test score distributions and 2 levels (15%/none other than item of interest) of proportion of biased items in test were crossed to give 32 experimental conditions. In each of these conditions, both uniform and nonuniform DIF were simulated with sizes of DIF fixed at one of 4 levels: an area value of .2, .4, .6, or .8. Twenty replications of each condition were carried out. Separate ANOVA's - one for uniform and one for nonuniform DIF) were carried out in which the dependent variable was the number of times out of the 20 replications that an item containing DIF was flagged by each procedure. The independent variables were all the five factors identified earlier.

For uniform DIF, both test length and shape of score distribution showed no significant effects for both indices. Model-data fit had no effect on the LR procedure's power while the MH procedure's power appeared to be affected by model-data fit. The mean success percentage difference between the good-fit (2PL) model and the poor-fit (3PL) model was only 2% (i.e. 73% for the 3PL and 71% for the 2PL) yet this difference turned out to be significant.

The proportion of biased items in the test showed no effect on the power of the MH procedure. The LR procedure's power, on the other hand, was affected by the proportion of biased items in the test. The detection rate of the LR procedure rose from 70% to 76% when the proportion of biased items dropped from 15% to none other than the item under study. Sample size displayed a strong effect on the power of both procedures. When sample size increased from 250 to 500, the detection rate for both procedures increased by about 15%. Also, items with small DIF size (e.g. area value of 0.2) had low detection rates (about 30%) while items with large DIF size (e.g. area value of 0.8) had high detection rates (about 95%) for both procedures.

For the nonuniform DIF, shape of score distribution and proportion of biased items in test displayed no effects on the powers of both procedures. The MH procedure's power was not affected by test length. The LR procedure's power increased from 57% to 62% when test length increased from 40 to 80 items. Both procedures were sensitive to model-data fit. For the LR procedure, the detection rate was 15% lower in the poor-fit (3PL) data than in the good-fit data. For the MH procedure, the detection rate was 7% lower in the poor-fit data than in the good-fit data. When sample size increased from 250 to 500 per group, the LR and MH detection rates increased by 20% and 10% respectively. For types of items, the LR procedure had its lowest detection rate (about 37%) for items with medium difficulty and low discrimination, and its highest rate (about 78%) for items with medium difficulty. For

items with low difficulty, the MH detection rate was about 15% lower than that for the LR procedure. For items with high difficulty, the detection rates for both procedures were almost the same (78% for LR and 80% for MH). The MH procedure was equally effective in detecting nonuniform bias with differences in the difficulty parameters of the items for the two groups. However, when bias was purely in terms of difference in the item discrimination parameter in the two groups, the MH procedure was unable to detect such items.

A study to evaluate the validity of some non-IRT bias detection procedures was undertaken by Shepard, Camilli and Williams (1985). One of the main assumptions made in this study was that bias detection procedures based on the IRT model always produce reliable results because of their underlying theoretical soundness. Thus, in the study, non-IRT procedures are referred to as "Approximation Techniques". As a result of this underlying assumption, the study's main focus became the validation of the so-called approximation techniques in terms of their ability to replicate results obtained from IRT bias detection techniques. The study was in two parts. The first part of the study was an EDS study which used data from the High School and Beyond (HSB) data file from the National Center for Educational Statistics. The HSB data used was for 28,000 high school seniors. The study sample extracted from the HSB sample was made up of 3,377 blacks and 17,928 whites (excluding Hispanics) who responded to the HSB math test in the spring of 1980. The so called "approximation

techniques" used in this study were the Transformed Item Difficulty (TID) procedure, the Full Chi-Square (χ^2_{full}) procedure and the Pseudo IRT (IRT-Z) procedure. The IRT procedures used were the following 3PL IRT procedures: the Sum of Square (SOS) 1, 2, 3, and 4 procedures, the Signed Area (SA), the Unsigned Area (UA) and the IRT Chi-Square (IRT-χ^2) procedures.

In the first study, the "true" state of bias in the test data was established by a crossvalidation IRT criterion study which used bias results obtained by applying the SOS2 and the SOS4 procedures to the math test data for two pairs of black and white comparison samples of equal sample size (N=1500 each). Items flagged by both procedures, in both pairs of samples, formed the bias criterion set of items. The validity of the other procedures (especially the "approximation" procedures) was then judged by each procedure's ability to replicate the result of the cross validation study. In the validation study, the samples used were 1000 whites and 300 blacks obtained by randomly sampling from the 17,928 white and 3,377 black samples. For procedures with no established test of significance, baseline studies were used to establish their cut-scores.

The second study was a Simulated Data Study (SDS) in which data were simulated for a 54 item test for a white sample of 1000 subjects and a black sample of 300 subjects using the 3PL IRT model. The mean ability for the white and black samples was fixed at 0.8 and 0.0 respectively, with the variance for both samples equal to 1.0. The 54 items were generated using three levels (-1.0,

0.0 and 0.5) of difficulty (b) parameters, three levels (0.5, 1.0 and 1.5) of discrimination (a) parameters and a fixed guessing (c=0.25) parameter for each item. In this study, data for 18 of the 54 items were generated biased in the black sample by altering the b values for the items. The alteration of 9 of the 18 items was obtained by increasing their b values by 0.20 while the b values for the other 9 items were increased by 0.35. Baseline studies were again used to determine cut-scores for procedures with no established test of significance. A back-up study using matched distribution samples of 284 whites and 284 blacks was also undertaken. A modification of the original Full Chi-Square procedure - in the form of a two-stage Chi-Square procedure - was also introduced into the study.

In the first study (i.e. the EDS) the following results were obtained for the unsigned indices: the unsigned IRT-Z identified 6 out of the 10 criterion biased items with 1 false-positive identification; the unsigned χ^2_{full} identified 5 out of 10 criterion biased items with 3 false-positive; the matched distribution χ^2_{full} identified 4 out of 10 items plus 1 false-positive; and the two-step χ^2_{full} identified 7 out of 10 criterion items with 5 false-positives. For the signed indices, the results were: 7 out of 10 criterion items plus no false-positives for the signed IRT-Z; the signed TID identified 4 out of 10 criterion items plus 8 false-positives in the unmatched samples and 5 out 10 criterion items plus 2 false-positives in the matched samples. The signed χ^2_{full} identified 8 out of 10 criterion items plus 1 false-positive in the

unmatched samples and 4 out of 10 criterion items plus no false-positives in the matched samples. The two-stage signed χ^2_{ms} identified 6 out of 10 criterion items plus 3 false-positives. The percentage agreement between the results obtained for the different indices and the crossvalidation IRT results were also determined: The unsigned IRT-Z, unsigned χ^2, matched distribution χ^2 and the two-stage χ^2 had 83%, 72%, 76% and 72% agreements respectively with the crossvalidation IRT. The signed IRT-Z, signed TID, signed matched-distribution TID, signed χ^2, matched-distribution χ^2 and the signed two-stage χ^2 had 90%, 52%, 76%, 90%, 79% and 76% agreements respectively with the crossvalidation IRT result.

For the second study (i.e. the SDS) Pearson Correlation Coefficients between the true state of bias (as generated) in the data and the results obtained for each of the "approximation" procedures were computed. The unsigned two-stage χ^2, unsigned χ^2, unsigned matched-distribution χ^2, and the IRT-Z results correlated .58, .51, .36, and .25 respectively with the true biased state in the data. The signed IRT-Z, signed two-stage χ_2, signed χ^2, signed matched-distribution χ^2, signed matched-distribution TID, and the signed TID results correlated .62, .61, .59, .52, .52 and .46 respectively with the true biased state in the data. In this study also, the correlation coefficients between the true state of bias in the data and the results obtained by each of the IRT procedures were computed: For the unsigned IRT procedures, the SOS1, IRT χ^2, and SOS2 correlated .64, .62, and .58 respectively with the true biased state in the data. For the signed IRT procedures, the SOS3

and SOS4 correlated .64 and .61 respectively with the true state of bias in the data.

Seven DIF detection procedures - the transformed Item difficulty - major axis (TID - MA), the transformed Item difficulty - 45° line (TID - 45°), the one-parameter ICC - fit statistic (ICC-1F), the one-parameter ICC - difference in item easiness (ICC-1E), the three-parameter Area Between (ICC-3), the full Chi-Square with five ability intervals (CHI-5), and the full Chi-Square with multiple ability intervals (CHI-N) - were examined for their effectiveness in a simulated data study by Rudner, Getson and Knight (1980). The procedures' effectiveness was studied under 112 different experimental conditions obtained by crossing 7 test length levels (20, 30, 40, ..., 80 items) with 4 levels of bias in item discrimination (a) values and 4 levels of bias in item difficulty (b) values. The levels of bias in item discrimination values were obtained by setting the variance of the groups' differences in the "a" parameters ($\sigma\Delta a_i$), at 0, 0.4, 0.8 and 1.2 respectively. In a similar way, the levels of bias in item difficulty values were obtained by setting the variance of the groups' differences in the b parameter ($\sigma\Delta b_i$) at 0, 0.5, 1.0, and 1.5 respectively. Two groups, with 1200 examinees in each group, were used. The mean abilities for the groups were fixed at -0.5 and 0.5 respectively. Each group's variance was fixed at 1.0. Data generation was based on the 3PL IRT model.

With the exception of the CHI-N procedure, the results from all the other procedures displayed only a slight increase in

average correlation with the true bias state of the data as the test length increased. There was no significant increase in bias detection rate with respect to increase in test length for any of the procedures. The largest increase in correlation was for the ICC-3 procedure, between test lengths of 20 and 30 items, where the correlation with generated bias jumped from .67 to .81 respectively. Over all the 5600 items generated in the different conditions of the study, the correlation between the true biased state in the data and each of the procedures were .80, .73, .68, .63, .61, .60, and .55 for the ICC-3, CHI-5, TID-45°, CHI-N, TID-MA, ICC-1E and ICC-1F respectively.

Results from the ICC-3, TID-45 and ICC-1E showed an increase in correlation with true bias state as the variance of the difference in the groups' b parameters increased. The ICC-1F showed a steady decrease in correlation with true biased state as $\sigma\Delta b_i$ increased. For the two Chi-Square procedures, there was a decrease in correlation with the true biased state as $\sigma\Delta b_i$ increased until the extreme condition when $\sigma\Delta b_i = 1.5$. At this level the correlation of the Chi-Square procedures with the true biased state increased (i.e. from .66 and .57 at $\sigma\Delta b_i = 1.0$ to .79 and .67 at $\sigma\Delta b_i = 1.5$ for Chi-5 and Chi-N respectively). At $\sigma\Delta b_i = 0$ (i.e. for the nonuniform bias condition), the CHI-5, ICC-3, ICC-1F, CHI-N, TID-MA, TID-45° and the ICC-1E had correlations of .81, .67, .66, .65, .52, .45, and .34 respectively with the generated biased state. The correlation between the true biased state and the results of each of the

procedures tended to decrease with increase in $\sigma\Delta a_i$ except those for the ICC-1F, which displayed a steady increase, and the ICC-3, which was fairly stable, as $\sigma\Delta a_i$ increased. At $\sigma\Delta a_i = 0$, all the procedures, except the ICC-1F, had correlations with generated biased state, ranging between .76 and .86. The ICC-1F correlated only .50 with the generated biased state at $\sigma\Delta a_i = 0$.

When the results obtained for the different procedures were correlated among themselves, the two highest correlations of .87 and .85 were obtained between the results for CHI-5 and CHI-N, and the results for CHI-5 and TID-45° respectively. The correlations between the ICC-1F and each of the others were relatively low; ranging from .40 with ICC-3, to .29 with TID-MA.

Merz and Grossen (1979) examined the effectiveness of six DIF procedures - the Transformed Item Difficulty (TID), the Point Biserial Correlation (PBIS), the Chi-Square (CHI), the Factor Analysis (FACTOR), the one-parameter ICC-Area Between (BICAL) and the Three-parameter-Area Between (PARAM) procedures. The study used simulated three-parameter IRT model data for a 60 item test. Discrimination (a) and pseudo-guessing (c) parameters were kept fixed at a a=1.0 and c=0.05 respectively for all items. Sample sizes for the two study groups were also kept fixed at 1,000 examinees per group. Only the difficulty (b) parameters were manipulated between groups to create the required amounts of DIF in the selected items. Two difficulty levels of the test were used: In one, they maintained 60% of the items correct while in the second they allowed 80% of the items to be correct. How these test

difficulty levels translate into item difficulty (b) parameters for the test items was not mentioned. Three states of bias (0%,10% and 20% of the items biased) were built into each of the test difficulty levels. For each of the six resulting experimental conditions (2 levels of test difficulty and 3 states of bias), two major comparisons were made: In one, the degree to which each procedure's results correlated with the generated bias state was assessed. In the second, the accuracy with which each procedure flagged the generated biased items was estimated.

For the data set with 60% test difficulty level, when 10% of the items were biased, the correlation between the true (generated) bias state and results obtained from the DIF procedures were: .96 for TID, PARAM and BICAL, .86 for CHI, .10 for FACTOR and .07 for PBIS. When 20% of the items were biased the correlations became .98 for BICAL, .97 for TID, .95 for FACTOR, .94 for PARAM, .91 for CHI and .36 for PBIS. On the analyses for accuracy of identifying generated biased items, the results from the 60% test difficulty data were 83(4) for PARAM, 100(4) for BICAL, 33(7) for PBIS, 83(4) for TID, 83(2) for CHI, and 0(7) for FACTOR - where the number outside the bracket gives the percentage of correct identification and the number inside the bracket gives the number of false-positive identifications for the 10%-of-items-biased condition. For the 60% test difficulty and 20% of items biased, the results were 83(6) for PARAM, 92(4) for BICAL, 41(4) for PBIS, 92(7) for TID, 83(4) for CHI and 92(5) for FACTOR.

When the data sets with 80% test difficulty level and 10% of the items biased were analysed, the correlations obtained between the results from the DIF procedures and the generated state of bias were .95 for TID and BICAL, .92 for FACTOR, .87 for PARAM, .75 for CHI and .54 for PBIS. With 20% of the items biased, the correlations became .96 for TID, .95 for FACTOR, .75 for PARAM and CHI, .61 for PBIS and .20 for BICAL. On the accuracy of identifying generated biased items, the results from the 80% test difficulty and 10% of items biased condition were 50(6) for PARAM, 83(6) for BICAL, 67(6) for PBIS, 100(4) for TID, 33(2) for CHI and 83(5) for FACTOR. When the number of biased items was increased form 10% to 20%, the results became 69(7) for PARAM, 46(18) for BICAL, 62(9) for PBIS, 85(4) for TID, 23(3) for CHI and 92(5) for FACTOR - with numbers in and outside the bracket as defined in the previous paragraph.

In the four simulated data studies presented above, the study by Rogers and Swaminathan (1990) is in many ways different from the other three. The Rogers and Swaminathan study compared only two DIF procedures each of which was proposed after the other three studies had been done: The Mantel-Haenszel procedure was proposed by Holland and Thayer (1986), and the Logistic Regression procedure was proposed by Rogers and Swaminathan (1989). Thus there is no way by which the results from Rogers and Swaminathan (1990) could be compared with those from the other three studies. In summary, however, the main findings of Rogers and Swaminathan (1990) were I) the test statistics of both the LR and the MH procedures had their

expected distributions (i.e. they are both distributed as Chi-Squares with 1 degree of freedom for the MH procedure and with 2 degrees of freedom for the LR procedure) under nearly all the conditions examined. They observed that the LR test statistic may not have the expected distribution for items with high difficulty (b) and discrimination (a) parameters. II) The LR procedure was more powerful in detecting nonuniform DIF. III) Both the LR and MH procedures were equally powerful in detecting Uniform DIF.

Findings from both the empirical and simulated data studies were consistent in the Shepard, Camilli and Williams (1985) study. Among the approximation DIF procedures, they obtained the best result with the Pseudo IRT procedure. The full Chi-Square procedure was close in accuracy to the Pseudo IRT procedure. The Transformed Item Difficulty (TID) procedure was found inadequate. However, a modified form of the TID, in which the p-value differences were regressed on the item point biserial, was found to be as accurate as the Chi-Square procedure.

Rudner, Getson and Knight (1980) found the three-parameter ICC-Area Between (ICC-3) and the Chi-square, with five ability intervals, procedures accurate under all investigated conditions. The TID procedure was found to perform well on items with uniform DIF. On items with nonuniform DIF, the TID performed poorly. These researchers also found that increasing the number of ability intervals reduced the power of the Chi-square procedure. The one-parameter ICC-fit statistic and - Difference in Item Easiness

procedures consistently returned poor correlations with the generated biased states of the data.

Merz and Grossen (1979) found the TID procedure better than any of the other five procedures in their study. The three-parameter and one-parameter Area Between procedures where next to the TID procedure. The Factor Analytic procedure came next, in the hierarchy of power, to the Area Between procedures. The Chi-square procedure was less powerful than the Factor Analytic procedure although it was far better than the Point Biserial procedure which was found to function erratically.

One study that cannot be strictly classified as either a simulated or an empirical study is the study by Spray (1989). In this two part study, Spray examined the behaviour of three DIF procedures: The Standardized Difference in Proportion-Correct (STD) by Dorans and Kulick (1986), the Mantel-Haenszel Common Odds-ratio (MH) by Holland and Thayer (1986), and the Root Mean Weighted Squared Difference in Proportion-Correct (RMWSD) by Dorans and Kulick (1986). Samples were drawn from a population of examinees who had taken the three forms (test A, test B and test C) of the ACT Assessment Program (AAP) test. For each test form a base (B) and focal (F) samples were drawn from the population of examinees. Sample sizes for all B and F groups were fixed at 2,000 each. Test A and test B had white base and black focal groups. Test C had a male base group and a female focal group. The total sample (both base and focal grouped together) was used to estimate item parameters for each of the test forms.

In the first part of the study, Spray defined the theoretical asymptotic distribution form of each of the three DIF procedures. Using these theoretical distributions, she defined an arbitrary set of criteria by which item bias was judged. The asymptotic theoretical DIF statistic for each of the three procedures was then computed for each test item. The bias criteria were then applied to judge whether an item displayed DIF or not. For each test form, the bias result, obtained by using this theoretical form of the DIF procedure, was then used as the true bias states of the test items with respect to the population subgroups used.

In the second part of the study, the item parameters obtained earlier for items in each test form were used in simulating test data for nine sample size arrangements: Six arrangements had an F to B ratio of 1:1 (2000/2000, 1000/1000, 500/500, 250/250, 100/100, 50/50) and three arrangements had an F to B ratio of 1:10 (200/2000, 100/1000, 50/500). Data generation for each test form and for each of the nine sample size arrangements was undertaken and replicated a hundred times. These generated data were then used to compute the value of the index, for each of the ordinary forms of the three DIF procedures, for each item. The average value across items for each index was computed and compared with its corresponding asymptotic counterpart. Both the STD and RMWSD were found to overestimate their respective asymptotic averages at all sample size levels for each of the three test forms.

Spray also considered the DIF-item identification power of each of the DIF procedures at the individual item level. For each

item, the number of times (out of a hundred replication) it was flagged by a procedure was determined and recorded in percentage form. Judgement on items was based on the degree to which the result for an item reflected the result of the asymptotic distribution study for the same items in the same test form. In Test A, for example, items 5, 7, and 19 were identified as displaying DIF in the asymptotic study. The MH procedure flagged these items 46%, 34% and 92% of the times, respectively, for the 250/250 samples. These rates changed to 57%, 46% and 76% respectively when the samples became 100/100. For the same items, the STD procedure's rates were 74%, 56% and 52% respectively for samples 250/250. When the samples were reduced to 100/100 the rates became 56%, 73% and 56% respectively for the STD procedure.

In the asymptotic study, items 6, 12 and 20 of test A were clearly identified as displaying no DIF by the three procedures. In the second part of the study (i.e. using the normal forms of the indices). The MH procedure flagged these items as biased 0%, 0% and 0% respectively for samples 250/250. These rates changed to 10%, 12% and 13% respectively for samples 100/100. The STD procedure identified these items 0%, 0% and 2% of the time respectively for samples 250/250. When the samples became 100/100, the rates changed to 29%, 29% and 21% respectively for the STD procedure. For the same items, the RMWSD procedure's rates were 100%, 100% and 100% respectively for each of the two sample sizes.

In summary, Spray noted that the MH procedure tended to slightly underestimate its asymptotic results for samples 2000/2000

and overestimated it as the sample sizes decrease. The STD and RMWSD procedures overestimated their asymptotic results for all sample sizes. The MH and STD procedures flagged truly biased items at similar rates for samples of size 250 or greater. However, the STD procedure tended to misclassify more items as biased in favour of the focal group than the MH procedure for smaller samples. The RMWSD, on the other hand, identified almost all items as displaying DIF even for samples as large as 500 to 1000.

Because of its tendency to classify items as displaying DIF when, in fact, they do not (i.e. it displays an unacceptable Type I error rate), Spray could not recommend the use of the RMWSD procedure. The other two procedures, the MH and the STD procedures, she classified as viable, useful candidates for use in DIF studies. She, however, draws attention to the unexpected finding of the tendency of the MH procedure to yield unbiased results for moderate sample sizes (e.g. 500 to 250) and to underestimate its asymptotic value for large samples while overestimating it for small samples. The STD procedure, on the other hand, consistently overestimated its asymptotic results irrespective of sample size. Since the MH procedure gives "nice unbiased-like" results for moderate sample sizes and moderate test lengths, Spray feels this may give it enough advantage for its use over the STD procedure.

Empirical Data Studies

Schulz, Perlman and Wright (1989) empirically compared the Rasch and the Mantel-Haenszel (MH) procedures for detecting DIF.

This study was undertaken mainly to verify claims of similarities, between the procedures, which is a result of the fact that both procedures were based on the one-parameter unidimensional model. The study examined the magnitude, empirical error variance, total variance and reliability of the indices generated by the procedures under selected experimental conditions. Two forms of the Rasch procedure and three forms of the MH procedure were used. In addition, the MH forms were used with the studied item included or excluded from the computation of total test score used for the measure of ability for the MH procedure. The number of ability intervals used for the MH procedure was either seven or all score levels where each obtained test score becomes an ability level.

Data for this study were obtained from the records of 60,000 examinees who had responded to a 46 item minimum proficiency skill test. The 60,000 comprised of 30,000 males and 30,000 females, 19,980 of the examinees belonged to one racial group (labelled race 1) and 20,040 belonged to another racial group (labelled race 2). Means for males and females were similar at 32.0 and 32.1 respectively. Means for the two racial groups were dissimilar at 35.5 and 30.0 for race 1 and race 2 respectively. Both sex and race contrasts were used in the study. A 1:1 majority to minority samples ratio was used in all cases in the study. Three sample sizes (100, 300 and 1000 per group) were used for the sex contrast. For the race contrast two sample sizes (200 and 666 per group) were used. Thirty samples were drawn randomly (without replacement) for each contrast (Sex or Race) in each sample size category. This

arrangement guaranteed that when the procedures were applied to the test results, for each index and combination of contrast and sample size, there would be 30 randomly equivalent, but independent estimates of DIF for each of the 46 items in the test. Thus for each procedure and in each experimental condition, the set of estimates formed a matrix of 46 items (rows) by 30 index-values (columns).

The error variance (E) of an item was determined by estimating the variance within the item but across the 30 columns - i.e. the row variance of the matrix. Total variance (T) was estimated by the variance of all 46 x 30 (i.e. 1380) values in the matrix. The net bias for an index was the grand mean of all 1380 values in the matrix. This grand mean is expected to be zero if item bias is not confounded with group differences in test performance. The reliability (R) of an index was estimated by the ratio of true (total - error) variance to total variance. Thus R=(T-E)/T.

The pooled within item, but across replication, error variance (E), of the standardized forms of the bias indices, was expected to be 1.0 in each case. In all the conditions, except the one for the sex contrast with sample 1000/1000, Rasch error variance was closer to 1.0 than the corresponding MH error variance. For the random sample groups, the total variance (T) was less than 1.0. This matches the expected error variance when there is no DIF - i.e. when the groups are random samples, the total variance should be equal to the error variance. This was almost exactly the case in this study. When groups are not randomly equivalent, the total

variance may not be equal to the error variance in which case, the possibility of bias exists. In such a situation, the total variance (T) is expected to exceed the error variance (E). In the study, total variance exceeded error variance for all the sex and race contrasts. However, the total variance for the MH indices was noticeably more conservative than for the Rasch counterparts. For the sex contrasts, the total variance for the Rasch procedures ranged from 1.37 to 5.66 while those for the MH ranged from 1.00 to 5.60. For the race contrasts the Rasch total variance ranged from 1.41 to 2.56 while those for the MH ranged from 1.24 to 2.68. For small samples (e.g. 100/100) Rasch total variance was 1.37 while MH total variance ranged between 1.00 to 1.10. For larger samples (e.g. 1000/1000) Rasch and MH total variances were similar, 5.66 for Rasch while those for MH ranged between 5.42 and 5.60.

Comparing the reliabilities of the procedures, Schulz et al. (1989) found that the Rasch reliabilities exceeded those for MH at the different sample size levels. The advantage of the Rasch over the MH, in reliability, was greatest when the sample sizes were small. For example, for sample 100/100, the Rasch reliability coefficient was .33 while those for the different MH procedures ranged between .26 and .30. In the race contrasts, the Rasch reliabilities were about equal to those for the MH procedure that includes the studied item when total test score is computed. The MH procedure that excludes the studied item in total test score computation yielded slightly larger reliability coefficients than the corresponding Rasch results. For example, when the studied item

was included in the computation of total test score, the MH with each obtained score used as ability interval and the MH with seven ability intervals gave reliabilities of .36 and .37 respectively while the corresponding Rasch reliability was .37 for samples 200/200 in the race contrasts. In the same race contrast, with samples 666/666, the corresponding reliabilities for the two MH's became .63 and .64 while that for Rasch became .63. When the studied item was excluded, in the computation of total test score, the 200/200 sample in the race contrast gave reliabilities of .39 and .39 for the two MH's (all ability intervals and seven ability intervals respectively). These became .66 and .66 respectively when the samples increased to 666/666.

The Rasch and all the variations of the MH procedures did not confound bias with differences in test performance when the groups were randomly equivalent. When the groups were not randomly equivalent, only one variation of the MH procedure (i.e. the MH variation with all score levels and studied item included in total test score computation) was equal to the Rasch in terms of not confounding bias with differences in groups' test performance. The MH variation that did not confound bias with differences in groups' performance was then compared with the Rasch in terms of detecting biased items. The two procedures displayed disagreements on eight of the items. For each of these items one of the procedures flagged it as displaying DIF while the other did not. Each of the eight items displayed strong 'misfit' with the Rasch model. The

researchers therefore concluded that these items may possibly be violating some of the assumptions of the MH procedure.

In conclusion, the researchers found no empirical basis for arguing the superiority of the MH procedure over the Rasch procedure based on their findings on a) confounding bias with between-groups differences in achievement, and b) statistical power of procedure defined in terms of reliability.

Hambleton and Rogers (1988) compared the IRT Area and the Mantel-Haenszel (MH) procedures. The study focused on three major areas: agreement between methods in identifying DIF, possible reasons for disagreement between methods, when they occur, and behaviour of DIF statistics when the ability distributions of the two groups are considerably different. The study utilized the results of 23,000 students who had taken the 1982 administration of the New Mexico High School Proficiency Examination (NMHSPE). The examination was a 150 item test that assesses "Life skills". Of the 23,000 students, 8,000 were Anglo-Americans while 2,600 were Native-Americans. These two subgroups were the study groups in the experiment. Four samples - two from each subgroup - each of 1,000 students were randomly selected. To establish a 'cutoff' value for the IRT Area index, two randomly equivalent samples were drawn from the Native-American subgroup. The largest area index value obtained with these two randomly equivalent Native-American samples was accepted as the largest value of the index that is likely to occur by chance. The study used only 75 of the 150 items in the NMHSPE.

Two independent DIF analyses were undertaken using the two Anglo-American and the two Native-American samples. The second comparison was designed to aid in establishing the degree of consistency of each of the two procedures.

To examine the effect of discrepancy in score distribution on the area index, the ability interval, over which the area index was calculated, was modified to reflect the region in the ability scale where most of the Native-Americans were located. For the MH index, the effect of score distribution differences was examined by way of a matched-group analysis. The matched-group analysis used a sample of 650 Native-Americans whose score distribution closely matched the score distribution of a sample of 650 Anglo-Americans.

In the consistency study, both methods performed relatively poorly. The area method flagged 20 of the 75 items in one comparison but not the other. The MH method flagged 15 of the 75 items in one comparison but not the other. The overall consistency for the area method was 73% while that for the MH method was 80%.

To compare the bias results obtained by the two procedures, only items consistently flagged by a procedure across the two comparisons were counted. In all, the two methods together flagged 16 of the 75 items. The area method consistently flagged 14 of the 16 items while the MH method consistently flagged nine of the 16 items. Of the nine items flagged by the MH method, seven were common to the area method. Thus two items were flagged by the MH method and not by the area method. Seven items were flagged by the area method and not by the MH method. Of the seven flagged by the

area method and not by the MH method, two were flagged in one of the two MH analyses. Thus a Type II error could have caused the problem with the MH method. Of the remaining five items, four were observed to display nonuniform DIF to which the MH procedure is not sensitive. The remaining item, which potentially displayed DIF against the Anglo-American group, had ICC's that were clearly different. This made it easily detectable by the area method. The MH procedure failed to flag this item because the region, on the ability scale, where the largest differences in probability of success were recorded, was a region where very few Anglo-Americans fell. Of the two items consistently flagged by the MH and not by the area method, one was flagged by the area method in one of the two analyses. Thus, the failure of the second area analysis to flag this item can be associated with a Type II error in the area method. The second item displayed uniform DIF but the ICC's were not markedly different for the two subgroups. Thus the area between the ICC's was smaller than the established 'cutoff'.

In the study with the modified ability range, the area method flagged 12 of the 14 items flagged in the original study. One of the two items, not flagged in this study, had ICC's that crossed within the modified ability range but did not diverge markedly within this new range. Other items, with crossed ICC's, which displayed large area values in the original study, tended to give relatively low area measures in the modified ability study. The rank-order correlation between the area method and the MH method

for the first original comparison was .32 while for the modified-ability the correlation was .48.

The MH procedure's matched samples results displayed very little change. Fourteen items, which included the nine in the original study, were flagged in the matched sample study. Four of the remaining five were flagged in one of the two original analyses. One additional item which was not flagged in the original study was flagged in this study.

A study to examine the reliability and validity of some IRT and non-IRT DIF detection procedures was undertaken by Skaggs and Lissitz (1988). They used samples from two independent populations to test the stability of the different indices. The IRT procedures used in this study were Area Between ICC's, Lord's Chi-Square, the SOS2 and the SOS4. The non-IRT procedures were the Full Chi-Square, the Mantel-Haenszel Chi-Square (MH CHI) and the Mantel-Haenszel Delta (MH Delta). The test data used in the study were from two different adminstrations of a curriculum-based general mathematics test for eight graders in a large suburban public school district in the U.S.A. Approximately 2,600 students took the test in its "field-testing" stage in the spring of 1986. In the spring of 1987, about 4,000 students took the first operational administration of the test. This study used 92 of the items in the test.

All DIF comparisons were made between samples of males and females. With respect to the eight graders in the study, the raw score distributions for males and females were identical. Males and females were compared on each item under the following conditions:

I) 1986 sample (n=650 females, 650 males); II) 1987 sample (n=2000 females, 2000 males); III) Reduced 1987 sample (a stratified sample to approximate the 1986 sample size and sex breakdown n=650 females, 650 males). Both the two- and three-parameter IRT models were used in each of the three conditions for the IRT DIF procedures. Stratified random samples (2,000 students in each sample) were used in baseline studies for each of the indices. In the baseline study also both the two- and three-parameter IRT models were used. No formal test of fit of data to the two- and three-parameter IRT models was undertaken. However, factor analysis on the results from the 1987 adminstration revealed one dominant factor accounting for between 15 and 17 percent of the total variance in each case. Each additional factor accounted for not more than 3 percent of the total variance.

Means for each IRT index were compared across the two- and three-parameter IRT models. The largest differences in mean values between the two- and three-parameter model results were 0.06 for the area method on the 1987 sample, 1.20 for Lord's Chi-Square on the reduced 1987 sample, 0.20 for SOS2 on both the 1986 sample and the full 1987 sample, and 0.73 for SOS4 on the full 1987 sample. Sample size was found to affect the means of the Chi-Square indices. The largest mean for each Chi-Square index was obtained for the full 1987 sample which was about 3 times the size of the other samples. The mean Lord's Chi-Square values for the two- and three-parameter model results (in the full 1987 sample) were 11.16 and 10.17 respectively while the highest mean for Lord's Chi-Square

in both the 1986 and reduced 1987 samples was 6.96. Except for the Lord's Chi-Square index, the mean baseline study value for each index was less than the mean values obtained in all the other analyses for each of the IRT indices. The Lord's Chi-Square index gave mean baseline values of 7.89 and 7.57 for the two- and three-parameter analyses respectively. These baseline means turned out to be larger than the means obtained for both the 1986 and reduced 1987 samples. Like the full 1987 samples, the baseline samples were also about three times the size of the 1986 and reduced 1987 samples. This may have caused the higher mean values for the baseline Lord's Chi-Square index over the means of the 1986 and reduced 1987 samples. The SOS indices also tended to produce larger means for larger sample sizes although this did not affect the baseline value as much as in the case of the Lord's Chi-Square index.

Spearman rank order correlations (r_s) were computed for each IRT index with itself across the different experimental conditions. The SOS4 gave the highest r_s values which ranged from .37 to .89. The SOS2 was next with r_s values ranging from .31 to .77. The area method gave r_s values ranging from .18 to .72 while Lord's Chi-Square was last in the group with r_s values ranging from .16 to .68. However, the r_s values for the different IRT indices computed across IRT models (i.e. the two- and three-parameter models) and across samples were lower than those computed across IRT models but within the sample. For the MH Chi-Square and the full Chi-Square procedures r_s values ranged from .15 to .55, when correlated with

themselves between the 1986 and reduced 1987 samples. Between the 1986 and full 1987 samples the MH - and Full Chi-Square gave correlations ranging from .24 to .69 with themselves.

Agreement between the DIF indices was evaluated in terms of intercorrelations between procedures in each of the experimental conditions. The correlation between the two signed indices (MH Delta and SOS4) were the highest and ranged from .84 to .90. These two indices, however, correlated close to zero with each of the unsigned indices. The correlation between MH Chi-Square and Full Chi-Square ranged from .57 to .84. The area method and Lord's Chi-Square correlated from .80 to .88. All the other indices correlated poorly ranging from .40 to .60.

Consistency with which items were flagged by each procedure, across experimental conditions, was also examined. There were six experimental conditions (two IRT models x three samples - 1986, reduced 1987 and full 1987 samples), for the IRT procedures and only three for the non-IRT procedures. The area method flagged no single item in all six conditions; only three items were consistently flagged in five of the conditions. Lord's Chi-Square consistently flagged 1 item in all six conditions. The SOS2 and SOS4 consistently flagged four and five items respectively in all six conditions. For the non-IRT procedures, the MH Delta, MH Chi-Square and Full Chi-Square consistently flagged five, six and four items respectively in the three conditions. There was no item flagged by all the procedures in all the conditions.

Skaggs and Lissitz (1988) observed that in their study, the correlation results indicated that reliability of each index declined when totally independent samples (i.e. samples from the 1986 and 1987 grade 8 student population) were used. In general, the sizes of the correlations obtained were quite small as measures of reliability. This casts a serious doubt on the indices' ability to identify DIF with any degree of consistency in realistic situations.

Seong and Subkoviak (1987) compared Linn and Harnisch Pseudo IRT(Z), Angoff Revised TID, and the Full Chi-Square techniques in terms of their performance in detecting item bias. The data used in this study were from a test (given to white and black examinees) in which some of the items were constructed biased in favour of blacks. Because in real life testing situations the number of minority group members is normally much smaller than the number or majority group members in the total examinee sample, this condition was built into its framework of the experiment.

The test used was a standard vocabulary test into which 10 "black slang" items were included. The test was made up of 50 multiple-choice items which included the 10 "black slang" items. The 10 "black slang" items were constructed by black experts and were classified as biased against whites purely on judgemental basis. The test was administered to an examinee sample containing 1,022 whites and 1,008 blacks. In the analysis for the study, all of the white examinees were used and only 300 black examinees were used.

For the analysis, both signed and unsigned indices for the three procedures were computed. An a priori bias index was devised which assigned (1) to each of the 10 black slang items and (0) to each of the other 40 items. How well a procedure recaptures the original bias in the data was measured in terms of the point biserial correlation coefficient between the a priori (0,1) codings of the test items and the index value for each item obtained for each procedure. An agreement statistic (i.e. percentage of items in agreement with a-priori state) was also computed. The researchers also examined the degree of agreement, among the different procedures, in terms of their Pearson correlation coefficients.

In terms of agreement with the a-priori bias state of the data, the IRT(Z) and the Full Chi-Square indices correlated .71 and .69 respectively (for their unsigned indices) and .76 and .80 (for their signed indices) with the a-priori bias state. The signed and unsigned Revised TID procedure correlated only .52 and .35 respectively with the a-priori bias state. The percentage agreement with the a-priori bias state in the data for the IRT(Z), the Full Chi-Square and the Revised TID were 92%, 92% and 76% (for unsigned measures); 98%, 100% and 94% (for signed measures) respectively. For the signed and unsigned, measures respectively, the correlation between IRT(Z) and the Full Chi-Square were .89 and .90; IRT(Z) and TID were .45 and .40; the Full Chi-Square and TID were .50 and .35.

From the results above, the researchers concluded that the Full Chi-Square index was slightly more accurate than the IRT(Z) index in detecting bias. The Revised TID was considerably worse

than any of the other two procedures in bias detection. There was high agreement in the results obtained by the IRT(Z) and the Full Chi-Square as evident in their high intercorrelations (.90 and .89).

Known Strengths and Wesknesses of Existing DIF Procedures

IRT Procedures

In situations where there are large enough samples and multiple-choice test items with nonuniform discrimination parameters, the only justifiable IRT DIF detection procedures are those based on the three-parameter IRT model. These procedures include: I) procedures based on the area between the ICC's of the studied groups (Rudner, 1977; Shepard et al., 1981; Linn et al., 1980, 1981), II) procedures based on the sum of squared differences in probability of correct response [$P_i(\theta)$] between the groups (Linn et al., 1980, 1981; Shepard et al., 1984), III) procedures based on a pseudo IRT model (Linn & Harnisch, 1981) IV) the Lord's Chi-Square procedure - even though this procedure ignores the pseudo-guessing (c) parameter (Lord, 1980)), V) procedures based on visual examination of ICC's (Linn et al., 1980, 1981) and VI) procedures based on empirical item response functions (Hulin et al., 1982, 1983; Bougon and Lissak, 1981).

The "Area Between ICC's" DIF procedures designed by Rudner and by Shepard et al. differ from each other only in the fact that Rudner's procedure uses a simple summation operation while the

procedure designed by Shepard et al. uses calculus in the form of an integral operation. Rudner's procedure can be equated to the procedure, designed by Linn et al., which uses the absolute values of the area measures that are summed. The two procedures will yield the same results in all cases. However, if the Linn et al. base-high and base-low version of the Area Between procedure is used on items displaying nonuniform DIF, the results will be different from those obtained by using Rudner's procedure for the same items. In any comparative study on DIF detection procedures, including one version of the area between procedures should be enough. In this study, an area between measure is used to fix the amount and direction of bias in the simulated biased items. It is believed that this will give an undue advantage to any Area Between procedure included in the study. For this reason, none of the Area Between procedures is included in the study.

Procedures based on the sum of squared differences in $P(\theta)$ - the probability that an individual with ability θ will answer an item correctly - have a basic similarity: Each of the procedures is built around values obtained by finding the differences between $P(\theta)$s for the studied groups at fixed ability (θ) values. Some of the indices in this group are signed (e.g. SOS3 and SOS4) while other are unsigned (e.g. SOS1, SOS2 and the Weighted Square Root of Sum of Square procedure proposed by Linn et al.). The Linn et al. index divides the ability scale into intervals and only $P(\theta)$s at the midpoints of these ability intervals are used in obtaining the index. The SOS procedures, on the other hand, use all the ability

(θ) values obtained by members of both studied groups. Some of the indices (e.g. Linn et al's., SOS2 and SOS4) are weighted by the inverse variance error of the P(θ) differences, so as to take care of large differences in the P(θ)'s which may be the result of poor estimation of parameters due to insufficient data at an ability level, (θ). Others (e.g. SOS1 and SOS3) are not weighted. The Linn et al. procedure also defines its index in square root form while the others do not. By using only the P(θ)'s at midpoints of ability intervals, the Linn et al. procedure does not utilize all the information in the data. The SOS procedures, on the other hand, use every obtained ability value and so utilize all the information in the data. However, if parameter estimations are poor, at any level of the ability scale, because of insufficient data, the SOS1 and SOS3 will be inappropriate because they have no built-in mechanism to take care of this problem. From the points raised so far, the only contenders, for place in a comparative study, among the Sum of Squared differences indices are the SOS2 and the SOS4. Since the SOS2 is unsigned while the SOS4 is signed, both of them were included in the study.

The Pseudo IRT model procedures proposed by Linn and Harnisch (1981) are "Lack of Fit" procedures designed to address the most practical problem encountered by prospective users of IRT models, large sample size requirement. In everyday testing situations, it is difficult, if not impossible, to meet the large sample size requirement of particularly the three-parameter IRT model. A procedure that can be used with small samples will be of immense

value in the everyday testing situations. Shepard et al., (1985) reported very favourably for the Pseudo IRT procedures, relative to other IRT procedures, when sample size is 300 or less. They also found the Pseudo IRT(Z) procedure more powerful in detecting items with DIF, than the full Chi-Square. Seong and Subkoviak (1987) found the full Chi-Square procedure more powerful than the IRT(Z) procedure. The Pseudo IRT procedures' claim to effectiveness with small samples, the mixed evaluation reports (as in the two studies mentioned above) they have received, and the fact that they have not been included in the same comparative study with more recent procedures like the logistic regression procedure, make them strong contenders for places in any contemporary comparative study of DIF procedures. However, in this study, neither the Pseudo IRT(Z) nor the Pseudo IRT (D) were included. The only reason for not including these indices in the study is one that has to do with the available finances and time at the disposal of the researcher. It is hoped that they will surely be included in future studies being considered by the researcher.

Not many comparative studies have included Lord's Chi-Square procedure. This is probably due to the mathematical sophistication of the procedure. Even though it utilizes two (difficulty (b) and discrimination (a)) out of the three item parameters in a three-parameter model, its theoretical soundness cannot be easily compromised. This index is one of those used in this study.

The visual procedure proposed by Linn et al., (1980, 1981) is quite an attractive procedure in terms of simplicity. All it

involves is plotting the ICC's (of the studied groups) and their confidence intervals on a common scale. However, the suggestion of using plus and minus twice the standard error, of each group's P(θ) values, to determine the confidence interval for that group's ICC cannot be objectively defended as the basis for judging DIF. This procedure is excluded from the study.

DIF detection procedures based on empirical item response functions (EIRF) have a weakness similar to those of the SOS1 and the SOS3. In the ability intervals created, it is possible to have some ability intervals into which very few subjects fall. If, for example, only five subjects in a particular group fall into interval g, and suppose four out of the five got the item correct, then the proportion of correct response for that group, in interval g, is 0.80. Suppose there is an ability interval K which is at a higher level than g; and suppose in the same group of subjects, 400 subjects fall into interval K. If 300 of the 400 subjects got the item correct, then the proportion of correct response for the group in interval K will be 0.75. Clearly there is a discrepancy in such results; the only explanation that could be advanced for such a situation is that of insufficient data in interval g. For this reason, the EIRF procedure needs some more refinement if it is to be considered a contender for a place in any serious comparative study on DIF procedures.

Non-IRT Procedures

Among the Chi-Square DIF procedures, the Scheuneman's (1979) procedure has been criticized on the grounds that the $\chi^2_{(correct)}$ index is not distributed as a Chi-Square (Baker, 1981; Marascuilo & Slaughter, 1981). This creates a problem in terms of its theoretical soundness. However, the weakness in Scheuneman's procedure was addressed by Camilli (1979) who developed the full Chi-Square procedure that utilizes both the correct and incorrect responses. The resulting index from the full chi-square procedure is known to be distributed as a Chi-Square. The full Chi-Square procedure has been featured in several comparative studies on DIF procedures. Its ratings from these comparative studies can best be described as mixed: Shepard et al. (1985) rated the full Chi-Square second to the Pseudo IRT(Z) which received the highest rating among the set of DIF procedures studied. Seong and Subkoviak (1987) gave a superior rating, over the Pseudo IRT, to the full Chi-Square procedure. Rudner, Getson and Knight (1980) placed the Chi-Square procedure on the same level as the one- and three-parameter area between procedures with respect to DIF detection power. Merz and Grossen (1979) found the Chi-Square procedure's power better than only the point biserial index which was described as behaving erratically. Despite the mixed evaluation reports on the Chi-Square procedure it still receives meaningful attention in test development situations. For these reasons, the full Chi-Square procedure is one of the procedures examined in this study.

The Mantel-Haenszel (MH) DIF procedures are among the high popularity DIF procedures used in the field today. Apart from the cost, small sample size and less complexity advantages these procedures have over other popular procedures (e.g. those based on the three-parameter IRT model), results from comparative studies have shown very strong agreements, in terms of DIF identification, between the MH procedures and other leading DIF detection procedures (see, for example, Hambleton & Rogers, 1988; Welch, Ackerman & Doolittle, 1987; Wright, 1986). Holland & Thayer (1986) highlighted some technical advantages the MH procedures have over other procedures, especially other procedures based on the Chi-Square distribution. The most important of these advantages is the fact that the MH procedure tests the hypothesis of no bias (H_0) against a specific alternative hypothesis (H_1). The main criticisms of the MH procedures are 1) insensitivity to nonuniform DIF (see Swaminathan & Rogers, 1989, 1990; Hambleton & Rogers, 1988), 2) the general large sample effect on the Chi-Square index and 3) like all other procedures that use contingency tables, the MH procedures are also charged with loss of information when the ability measure, which is basically a continuous measure, is transformed into discrete measures. The inability to detect nonuniform DIF can be appreciably taken-care-of by using "signed-forms" of the procedures. With respect to the large sample effect on the Chi-Square test of significance, there is not much that can be done. However, in most of the DIF detection studies in normal practical situations, the samples are not expected to be very large. It is

only with national and state examinations that large samples will be a problem. It is true that converting a continuous variable into a discrete variable will normally be accompanied by some loss of information. In any case, the loss of information can be minimized by using an optimum number of discrete intervals, making sure that there are several categories in the part of the underlying distribution where the score density is high (Shaw, Hoffman & Haviland, 1987). Both the MH Chi-Square and the MH Delta are included in this study.

A very recently proposed DIF procedure is the Logistic Regression (LR) procedure proposed by Swaminathan and Rogers (1989). The procedure uses a standard logistic regression model for predicting a dichotomous dependent variable from given independent variables. In the procedure, the probability of a correct response to an item is modelled to depend on both group membership and ability. The index is distributed as a Chi-Square with two degrees of freedom. In their studies, Swaminathan and Rogers (1989, 1990) found the LR procedure to have an equal amount of power, in detecting uniform DIF, as the MH procedure. On the detection of nonuniform DIF, the LR procedure was found to be much more powerful than the MH procedure. Being a Chi-Square test of significance, it is expected it will have the same weakness, with respect to large samples, like all Chi-Square procedures. To this end, both the LR and MH procedures can be described as having the same weakness. The LR procedure, on the other hand, has the advantages of a) treating the ability variable as a continuous variable, and b) being

sensitive to nonuniform DIF. So far, only the researchers who proposed the LR procedure have reported comparative studies in which the procedure was compared to the MH procedure (Swaminathan & Rogers, 1989, 1990). The validity of the claims made by Swaminathan and Rogers (1989, 1990) is worth testing in some other studies involving other DIF procedures. The LR procedure is therefore one of the procedures examined in the current study.

Spray and Carlson (1986, pp.13) drew attention to the weakness in using models, based on contingency tables, in analyses where at least one of the variables is a continuous variable. They also highlighted (pp.12-13) the conditions under which loglinear and logit models lead to the same exact results. Swaminathan and Rogers (1989) also mentioned the similarity between Mellenberg's (1982) loglinear DIF procedure and their logistic regression procedure which is built on a logit model. Because of the similarity between the loglinear and the logit models a decision was taken to include only the logistic regression procedure in this study.

DIF procedures based on transformed item difficulty (TID) have one very strong criticism: That item difficulty is confounded with both ability and item discrimination. This is a weakness in the theoretical framework of procedures based on transformed item difficulty. However, in some comparative studies, these indices are known to have performed fairly efficiently (Burrill, 1981; Rudner et al., 1980; Merz & Grossen, 1979) relative to the more theoretically sound procedures. In addition to these positive comparative study results, TID procedures are relatively low in

implementation cost, less sophisticated statistically, and relatively easy to interpret to the lay public. For these reasons, it was decided to include a modified form of the TID procedure in this study.

Dorans and Kulick (1986) developed a standardized procedure and its root mean weighted squared differences (RMWSD) procedure. In reporting the development of their DIF indices, they reported that their indices were criticized by Holland who demonstrated to them that their indices are biased because they retain all the sampling errors in their point estimate of success probabilities. Their reply to Holland's criticism was that the problem was a sample size related problem, and that the problem could be overcome by taking "Large" samples. They, however, failed to indicate how large the samples should be. Because of this, the two procedures by Dorans and Kulick - The Standardized Difference (STD) and its Root Mean Weighted Squared Difference (RMWSD) were not included in this study.

The Multivariate Factor Structure Analysis model was also used to develop DIF detection procedures (Lei & Skinner, 1983; Rock & Werts, 1979; Merz, 1976, 1973; Green & Draper, 1972). Merz (1976, 1973) and Green and Draper (1972) based their procedures on the exploratory factor analysis model while Lei and Skinner, (1983) and Rock and Werts (1979) based theirs on the confirmatory factor analysis model. Against certain other procedures for detecting DIF, procedures based on the factor analytic model have the advantage of addressing the underlying true abilities of the examinees (Rudner,

Getson, & Knight, 1980). There are however two major problems with these procedures: (i) factor analysis is best suited for situations where the underlying measurement structure is a continuous one rather than a dichotomous one (Hulin, Drasgow & Parson, 1983); (ii) factor analysis involves some decision making problems such as type of correlation matrix to use, type of factor analysis model to use, number of factors to extract, and type of rotational scheme to employ (Rudner, Getson, and Knight, 1980). All this decision making will clearly introduce subjectivity into the procedures. Rudner, Getson, and Knight, (1980) reported that in a comparative study undertaken by Rudner and Convey, the factor analytic procedure yielded unsatisfactory results. In the study by Merz and Grossen (1979) the factor analytic procedure's results were not consistent over the different conditions, when 10% of the items were biased it identified 0% and 83% respectively for the 60% and 80% Test difficulty conditions. Mainly because of the subjectivity introduced by the decision processes mentioned earlier, none of the factor analytic procedures was included in the study.

Item-test correlation has also been used to define DIF procedures. Stricker (1982) and Green and Draper (1972) developed DIF procedures based on item-test correlation. Stricker (1982) based his index on a partial correlation model while Green and Draper (1972) based their index on the point biserial correlation model. The main criticism of the point biserial procedure was presented in Angoff (1982). Angoff indicated that the point biserical index would find some items to display DIF against one

group even when the two studied groups are only randomly different. This is because the point biserial index is such that it would yield a significantly less than perfect correlation even between two highly similar groups. Angoff went on to report a study by Hunter in which Hunter clearly demonstrated that the point biserial DIF procedure would always indicate DIF as long as the items in the test do not have a uniform difficulty index irrespective of the fact that the items were unbiased. In comparative studies based on simulated data (Shepard, Camilli, & Williams, 1984; Ironson & Subkoviak, 1979; Merz & Grossen, 1979) the correlations between the true state of bias in the data and the bias results obtained by the point biserial procedure were very low. In comparative studies based on real data (Shepard, Camilli, & Williams, 1984; Ironson & Subkoviak, 1979) the correlations between the results obtained by other DIF procedures and the point biserial procedure were very low.

The procedure based on the partial correlation model received some better comments from Angoff (1982) than the point biserial model procedure. This was mainly due to the fact that the partial correlation DIF procedure automatically controls for ability differences that might exist between groups, by partialling out the effect of ability. However, Stricker (1982) indicated a weakness in the procedure. This weakness is in the procedure's inability to detect cases of nonuniform DIF. It has the tendency to label a nonuniform DIF item as an item with no DIF. This is because the bias at different ability levels (in the case of nonuniform bias)

will cancel out and so return a no DIF index for the item. So far, not much has been done in terms of comparative studies, involving the partial correlation procedure, that might establish this procedure as a contender among established DIF detection procedures. Both the partial correlation and the point biserial correlation DIF procedures were left out of the current study because of the weaknesses highlighted above.

Veale and Foreman (1983) proposed a DIF detection procedure based on the analysis of the distributions of incorrect responses, distractor analysis (DA). Rudner, Getson, and Knight, (1980) found the DA procedure appealing because it avoids the assumption (made by some other procedures) that total test scores are perfect indicators of true ability. They claimed that the approach will help save the practice of discarding an entire item in order to purify a test instrument, because it helps identify particular distractors that may be causing bias. Scheuneman (1982), on the other hand, expressed concern over the model's assumption that the overpull, for the reference group against which the other group's performance is being compared, is zero for all options. It is also possible that the pull index for a group is more a reflection of the group's mean ability rather than a function of culture, as assumed by the model. This could be the case because the different groups are not matched on ability. The data simulation process used in this study would not permit the DA index to be used: The DATAGEN program is used to generate only the 0,1 (incorrect, correct) type

of data. No provision is made for distractors. Thus the DF procedure was not used.

Hills (1989) attempted to summarize the results from comparative studies on DIF procedures. Findings from different studies were used to evaluate the different procedures in terms of their advantages and disadvantages. Based on these results, Hills went on to classify the DIF procedures in terms of those that are most useful to practitioners and those that are not. One concern with Hills' work had to do with his definition of the bases for comparison of the procedures. The studies upon which his evaluation was made were different in many ways. Each study dealt with a different set of DIF procedures, used different data sets, different populations, different sample sizes, and even different research procedures. Given these varied conditions of the studies, it is difficult to see how a common measure could be defined and used in evaluating the different DIF procedures. Results from such an evaluation would be much more valid when the DIF procedures, about which judgement is to be made, are studied under the same conditions.

So far, studies like Shepard, Camilli and Williams (1985); Rudner, Getson and Knight (1980), and Merz and Grossen (1979) compared a number of DIF procedures under uniform experimental conditions. The main problem with these studies is that they were done before the development of some of the currently prominent DIF procedures like the Mantel-Haenszel (MH) procedures (Holland & Thayer, 1986) and the Logistic Regression procedure (Swaminathan &

Rogers, 1989). Most of the other studies found in the literature dealt with two, three or, at most, four DIF procedures (e.g. Swaminathan & Rogers, 1990; Spray, 1989; Schultz et al., 1989; Hambleton & Rogers, 1988; Seong & Subkoviak, 1987; Wright, 1986). The study by Skaggs and Lissitz (1988) compared seven DIF procedures (4 IRT procedures and 3 non-IRT procedures) on empirical data obtained from real testing situations. In this study, however, the true state of bias in the data was not known. The only comparison was in terms of the degree of agreement between the DIF procedures in the study.

Like the studies by Shepard, Camilli and Williams (1985), Rudner, Getson and Knight, (1980), and Merz and Grossen (1979), this study compares a number of DIF procedures. In addition, it includes the most currently popular DIF procedures that were developed after 1985.

A total of eight DIF detection procedures are examined in this study. These include three IRT procedures and five non-IRT procedures. The eight procedures are described in the next chapter which gives a brief description of the study.

CHAPTER III

Description of Current Study

The eight DIF procedures used in this study will be described below. Procedures based on the IRT will be presented first, followed by the non-IRT procedures. The chapter will end with the presentation of the purpose of the study.

IRT DIF Procedures

From the item response theory standpoint, an item displays DIF if examinees of comparable ability but from different population subgroups do not have the same probability of success on the item. Effectively, this translates into a difference between the ICC's of the item, for the two subgroups, drawn on a common scale. From earlier sections of this chapter, it is evident that there are currently quite a number of procedures based on item response theory. However, for the different reasons given in an earlier portion of this chapter, only the SOS2, the SOS4 and the Lord's Chi-Square procedures are used in this study:

1) SOS2 (Sum of Squares 2).

This is an unsigned IRT index proposed by Shepard, Camilli and Williams (1984). To obtain this index, the studied item's success probability $P(\theta)$, at each obtained ability (θ) level in both the

Reference (or majority) group and the Focal (or minority) group, is determined for both Reference and Focal groups. The difference between the P(θ)'s for each group member is computed and squared. These are then summed over all ability (θ) values. This sum is divided by the product of the total number of examinees in both the reference and focal groups and the variance of the difference in the success probabilities. Mathematically, the index is given by:

$$SOS2 = 1/(n_R + n_F) \sum_{i=1}^{n_R+n_F} \frac{\{P_{iR}(\theta_j) - P_{iF}(\theta_j)\}^2}{\sigma^2_{P_{iR}(\theta) - P_{iF}(\theta)}}$$

Where:

n_R = number of examinees in the Reference group,

n_F = number of examinees in the Focal group,

j = subscript that counts all instances of θ for either group ($n_R + n_F$),

$P_i(\theta_j)$ = estimated probability of success on item i at the j^{th} instance of θ,

$\sigma^2_{P_{iR}(\theta) - P_{iF}(\theta)}$ = estimated variance error of the difference in estimated probabilities.

(Shepard et al., 1984)

2) <u>SOS4 (Sum of Squares 4)</u>

This index can be described as the signed form of the SOS2 proposed by Shepard, Camilli and Williams (1984). The only difference between the SOS2 and the SOS4 is that in the SOS2 the

P(θ) difference is squared before summing while in the SOS4, the P(θ) difference is multiplied by its absolute value before summing. The SOS4 index is given by:

$$SOS4 = 1/(n_R + n_F) \sum_{i=1}^{n_R+n_F} \frac{(P_{iR}(\theta_j) - P_{iF}(\theta_j))\,(|P_{iR}(\theta_j) - P_{iF}(\theta_j)|)}{\sigma^2_{P_{iR}(\theta)} - P_{iF}(\theta)}$$

where:

n_R, n_F, j, $P_i(\theta_j)$, and $\sigma_{P_{iR} - P_{iF}}$ are as defined for SOS2

(Shepard et al., 1984)

3) Lord's Chi-Square

This is an asymptotic significance test which simultaneously tests the equality of the discrimination (a) and difficulty (b) parameters estimated for the two studied groups. This procedure was proposed by Lord (1980). In this procedure, the a and b differences are weighted by various entries in the variance-covariance, matrix combined for the two groups. The index obtained is distributed as a Chi-Square with 2(two) degrees of freedom. For the studied item i, the index is given by:

$$\chi^2_i = (V_{i,j=1} - V_{i,j=2})' \; (\Sigma_{i,j=1} + \Sigma_{i,j=2})^{-1} (V_{i,j=1} - V_{i,j=2})$$

Where:

j = 1 for one subgroup and 2 for the other subgroup,

$V_{i,j}$ = Vector for a and b parameters for subgroup j, and

$\Sigma_{i,j}$ = Covariance matrix for vectors of a and b parameters for subgroup j.

(Lord, 1980)

NON-IRT DIF Procedures

In earlier discussions, the reasons for selecting the following five non-IRT DIF procedures for this study were given. The procedures are The Transformed Item Difficulty (TID), the Full Chi-Square (χ^2_{full}) also known as the Camilli Chi-Square, the Mantel-Haenszel Chi-Square (MH χ^2), the Mantel-Haenszel Delta (MH Delta) and the logistic regression (LR) procedures.

1) Transformed Item Difficulty (TID)

The TID procedure was first proposed by Angoff (1972). Since then a series of modifications to the original version have surfaced. The form used in this study is based on a modification suggested by Angoff (1982). This modification is expected to take care of the item discrimination effect on the index.

To obtain the index, the proportion of examinees answering an item correctly is computed for each of the two studied groups (the reference and focal groups). This proportion, p-value, is computed for all the items in the test. Each item's p-value is normalized for each group by converting it to a Z score, corresponding to the

$(1-p)^{th}$ percentile, which is obtained from the table of the normal curve. The normal deviate Z_i score so obtained is divided by the item's point biserial correlation coefficient (P_{bis}) to obtain Z_i'. The Z_i' value is then transformed into delta-values (Δ') by the linear transformation:

$$\Delta_i' = 4Z_i' + 13$$

The delta values from the two groups, for each item, are used as co-ordinates for a point which is plotted on a number plane where the vertical axis represents the delta values from one group while the horizontal axis represents the delta values from the other group. When all the points, defined by corresponding deltas for each item, have been plotted, the scatter-plot normally assumes an elliptical shape. A line, that minimizes the perpendicular deviations from the points, is then fitted onto the graph. Normally, this line defines the Major-axis of the ellipse. Mathematically, Angoff and Ford (1973), defined this major axis as:

$$Y = AX + B$$

where,

$$A = \frac{(S_y^2 - S_x^2) \pm \sqrt{((S_y^2 - S_x^2) + 4r_{xy}^2 S_x^2 S_y^2)}}{2r_{xy}S_xS_y}$$

and

$$B = M_y - AM_x$$

with

M_y = Mean of the deltas plotted on the y-axis

S^2_y = Variance of the deltas plotted on the y-axis

r_{tf} = Correlation coefficient between the deltas for the two groups

The DIF index is defined as the perpendicular distance (D) of each point from the major-axis. For item i, D_i is given by:

$$D_i = \frac{AX_i - Y_i + B}{\sqrt{(A^2 + 1)}}$$

(Angoff & Ford, 1973)

2) Full Chi-Square (X^2_{full})

In general, the chi-square procedure attempts to determine whether examinees of the same ability level have the same probability of a correct response for the studied item irrespective of group membership. This index was proposed by Camilli (1979). It is also called the Camilli Chi-Square.

To obtain the index, the total test score for each examinee in both the reference (R) and focal (F) groups is computed. The lowest and highest obtained test score in both groups, taken together, are determined. The range of scores between the lowest and the highest obtained scores is then divided into a fixed number of intervals (in this study six intervals were used). The number of examinees falling into each score interval, j, N_{jR} and N_{jF}, for the reference and focal groups respectively, is determined. Within interval j, the proportions, p_{jR} and p_{jF}, of examinees in the reference and focal groups respectively who answered the item correctly, are determined. Reference and focal groups are combined and the

proportion p_j, in the combined group, for interval j, students who answered the item correctly is determined. For the combined group, in each interval, $q_j = 1-p_j$ is computed. The index χ^2_{MH} for the studied item is then computed by:

$$\chi^2_{MH} = \sum_{j=1}^{j} [N_{jR} \frac{(p_{jR} - p_j)^2}{p_j q_j} + N_{jF} \frac{(p_{jF} - p_j)^2}{p_j q_j}]$$

with j(G-1) degrees of freedom

Where:

G = number of population subground
j = number of score intervals

(Camilli, 1979)

3) Mantel-Haenszel Chi-Square (MH χ^2)

Holland and Thayer (1986) made the following observations about the Full Chi-Square DIF procedure: i) That the Full Chi-Square procedure tests the hypothesis of equal proportion answering the studied item correctly, in each score interval for the groups under study, against a very general alternative hypothesis that is simply a negation of H_0 (the hypothesis of equal proportion); and ii) that only the hypothesis, H_0, is tested by the procedure, and produces no parametric measure of the amount of DIF exhibited by the studied item. To take care of these problems, Holland and Thayer (1986) proposed a procedure based on a method developed by Mantel and Haenszel (1959) for studying matched groups. This procedure, now known as the MANTEL-HAENSZEL CHI-SQUARE DIF

procedure, tests H_0 (the hypothesis of equal proportion) against a specific alternative hypothesis (H_1) given by

$$H_1: \; p_{Rj}/q_{Rj} = \alpha \, p_{Fj}/q_{Fj} \qquad j = 1,2,\ldots,k, \text{ and}$$

$$\alpha \neq 1$$

where,

R = One of the two studied groups labelled the "reference group",

F = the other studied group labelled the "focal group",

p = probability of answering the item correctly,

q = (1-P), the probability of answering the item incorrectly,

j = the j^{th} score interval (j=1,2,...,k),

α = parameter called the common odds-ratio across the k 2x2 tables -- an example of a 2x2 table for the j^{th} score interval is given below:

		SCORE ON STUDIED ITEM		
		1	0	TOTAL
GROUP	R	A_j	B_j	n_{Rj}
	F	C_j	D_j	n_{Fj}
	TOTAL	M_{1j}	M_{0j}	T_j

where,

R, F and j are as given above

A_j = number of examinees in R answering item correctly,

B_j = number of examinees in R answering item incorrectly,

C_j = number of examinees in F answering item correctly,

D_j = number of examinees in F answering item incorrectly,

n_{Rj} and n_{Fj} = number of examinees in Reference and Focal groups respectively at the j^{th} score interval,

m_{1j} and m_{0j} = number of examinees in the j^{th} interval answering the item correctly and incorrectly respectively, and

T_j = total number of examinees in the j^{th} interval.

The Mantel-Haenszel Chi-Square statistic is given by:

$$\chi^2_{MH} = [|\Sigma_j A_j - \Sigma_j E(A_j)| - \tfrac{1}{2}]^2 / \Sigma_j \mathrm{Var}(A_j)$$

With 1 degree of freedom.

where,

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

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

(Holland & Thayer, 1986)

An estimate of the common odds-ratio, α, across the 2x2 tables, is also provided by Mantel and Haenszel (1959). This is given by:

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

The scale of this odds-ratio extends from 0 to ∞. When $\alpha = 1$, this gives the null hypothesis, H_0, given by:

$$H_0: \; p_{Rj}/q_{Rj} = p_{Fj}/q_{Fj} \qquad j=1,2,\ldots,k$$

4) Mantel-Haenszel Delta (MH Delta or Δ_{MH})

Using the natural Logarithm of α_{MH}, results in a convenient symmetric scale in which 0 (the log of 1) becomes the value for the null hypothesis. Holland and Thayer (1986) further modified this log scale to fit the ETS "delta scale" which is a standardized scale. Their modification, Δ_{MH}, is given by:

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

5) Logistic Regression (LR)

This procedure, proposed by Swaminathan and Rogers (1989), is based on the standard logistic regression model for predicting a dichotomous dependent variable from given independent variables. In the procedure, the probability of a correct response to a differentially functioning item is modelled to depend on both group membership and ability. The model is given by:

$$P(U_{ij}=1) = \exp(\beta_{0j} + \beta_{1j}X_{ij})/[1 + \exp(\beta_{0j} + \beta_{1j}X_{ij})]$$

$$\text{with } i = 1,2,\ldots,n; \; j = 1,2$$

where,

U_{ij} = response of person i, in the j^{th} group, to the item,

β_{j1} = the intercept parameter for the j^{th} group,

β_{j1} = the slope parameter for the j^{th} group, and

X_{ij} = the ability of person i in group j.

When the logistic regression curves for the two groups are identical,

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

thus the item does not behave differentially.

When the two curves are not parallel, then

$$\beta_{01} = \beta_{02} \quad \text{but} \quad \beta_{11} \neq \beta_{12}$$

thus the item displays nonuniform DIF.

When the two curves are parallel and non-coincidental, then

$$\beta_{01} \neq \beta_{02} \quad \text{but} \quad \beta_{11} = \beta_{12}$$

thus the item displays uniform DIF.

(Swaminathan & Rogers, 1989)

Swaminathan and Rogers (1989) state that the parameters (β_{01}, β_{02}, β_{11}, and β_{12}) are asymptotically normally distributed. Thus, the statistic for testing the hypothesis of no bias, H_0, given by:

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

is a chi-square with two degrees of freedom.

The eight DIF detection procedures presented above are used in this study.

Purpose of Study

Half of the number of DIF procedures in this study produce indices that are distributed as chi-square. These are Lord's Chi-Square, Full Chi-Square, Mantel-Haenszel Chi-Square and the Logistic Regression procedures. It is a well known fact that the Chi-Square test statistic is inflated by sample size. This general weakness of the chi-square is not strong enough to override its strength when optimum samples are used. To this end, it becomes the responsibility of users, of the chi-square procedure to examine the effect of different sample sizes on it. With knowledge of the effect of sample size, on the chi-square index being used, the researchers would be able to meaningfully evaluate the result of the test statistic relative to the sample size used. For DIF procedures, like the four mentioned above, with indices distributed as chi-square, a study of their performance under different sample sizes will provide valuable knowledge for future users of the procedures. For the other four procedures, the SOS2, SOS4, TID and MH Delta, because their performance will be compared to those of the four procedures with indices distributed as chi-square, knowing how the non-chi-square indices perform within the same sample sizes

on which the chi-square indices are examined will greatly help the comparisons of procedures.

Test length (i.e. the number of items in a test) is known to have an effect on some DIF procedures. The length of the test affects the accuracy of total test score; longer tests produce more reliable total test scores than shorter tests. For procedures like the LR, MH chi-square and Full Chi-Square, where total test score may be used as a measure of ability, a longer test, which gives more reliable total test score, may also result in improved estimates of the indices of such procedures. In this study the performance of the different procedures when a short test and when a longer test are used is a condition on which comparison is made. It should be mentioned here that the concept of test length used here is modified to mean the number of items in a subtest that measures only one particular underlying trait. To this end, each of the tests used will be considered unidimensional.

All things being equal, the probability of success on an item strongly depends on the examinee's ability. Other factors that influence the examinee's probability of success on an item are the item discrimination and the item difficulty parameters. In item response theory, the two item parameters are represented by 'a' and 'b' the slope and intercept of the item's ICC, respectively (see Hambleton & Swaminathan, 1985; pp. 27-29). By the definition of DIF adopted in this study, an item is said to display DIF if examinees of comparable ability but from different subgroups within the target population do not have the same probability of success. This

definition clearly rules out difference in ability as a factor affecting DIF. It is also true that in real life testing situations, it is difficult (if not impossible) to find a test in which all the items have the same 'a' and the same 'b' parameters.

Estimation of the power, of each DIF procedure, to reproduce the true state of bias simulated in generated test data at different experimental conditions shall be examined.

To summarize, this study is divided into two parts: A null (no DIF) study and a known DIF study.

1. The Null (no bias) study:

   The objective of this study is to examine the effects of sample size, test length and selected arrangements of item discrimination values on the distribution of the eight indices.

   In addressing this objective, an attempt will be made to answer the following questions:

   - Are the cutoffs established at the .01, .05 and the .10 false-positive rates affected by sample size, test length, and/or arrangements of 'a' values?
   - For indices distributed as chi-square, how do the cutoffs at the .01 and .05 false-positive rates compare with the tabled values of chi-square?

2. The known DIF study:

   The objective in this study is to examine the effects of sample size, test length, discrimination arrangement, item difficulty level and type of DIF

(Uniform/nonuniform) on the power of the DIF procedures to reproduce the true state of bias simulated in the data and to compare their relative power in doing so.

To achieve this objective, answers to the following questions will be sought:

- Is the power to reproduce the true state of bias in the data affected by sample size, test length, discrimination arrangement, difficulty level, and type of DIF for each of the procedures?
- How is the false-positive rate of each procedure affected by sample size, test length, discrimination arrangement, difficulty level and type of DIF?
- What is the percentage of generated DIF items flagged by each procedure in each of the experimental conditions.

CHAPTER IV

Methodology

This study aims at (i) investigating the effect, if any, of sample size, test length and item discrimination arrangement on the distributions of the eight DIF detection procedures mentioned earlier; and (ii) investigating the effect, if any, of sample size, test length, item discrimination and item difficulty on the power of the eight DIF procedures in detecting items with DIF. This current chapter describes the methodology used in the study. The chapter is divided into sections which include: Data Collection, Assumptions Made in Study, Characteristics of Items, Characteristics of Samples, Data Simulation Model and Program, and Cutoffs and Data Analysis procedures.

Data Collection

Both parts of the study aim at examining certain characteristics of the eight DIF procedures identified previously. If meaningful evaluation of these characteristics is to be achieved, provision should be made, in the study, for the manipulation of factors that are believed to have effects on the characteristics of interest. For the characteristics of interest, such factors are sample size, test length, item discrimination and item difficulty. Manipulation of these factors can only be possible in a simulation study. This form of study permits the simulation of the conditions of the factors within which the researcher will want

to evaluate the performance of the different DIF detection procedures. A real data study will not provide similar amount of freedom to the researcher.

In the second part of the study, where the power to detect items with DIF is evaluated, unless the true state of DIF in the data is known, there is, as yet, no known means by which a valid examination of the power, of a given procedure, to detect DIF can be evaluated. There are some known studies (e.g. Shepard, Camilli & Williams, 1985; Subkoviak, Mack, Ironson & Craig, 1984) in which real data were used, and which claim to have examined the power of DIF indices to detect items with DIF. In such studies one of two methods is used to determine the "true state" of bias in the test data used. In one method, a crossvalidated result of DIF in the test data is obtained by using one of the three-parameter IRT DIF procedures. This result is then accepted as the true state of bias, in the test data, against which other DIF procedures' powers are evaluated. In the second method, items were constructed, by experts, to be biased against one of the studied groups. These items were then combined with other items, believed to be DIF free, to form a test. This test was then administered to the studied groups. The results from this adminstration of the test constituted the study data. It was then claimed, by the study, that the true state of bias in the data is defined in the results of the two sets of items (with respect to their bias nature, as spelled out above) combined to form the test. In the first method mentioned above, a major weakness is based on the fact that no DIF procedure currently

exists that is a hundred percent fool proof with respect to it's DIF detection performance. A big flaw therefore exists in accepting a DIF procedure's result, crossvalidated or otherwise, as a full definition of the true state of bias in a set of test data. For the second method, despite the common Gestalt fact that the whole is more than the sum of its parts, there is the established fact that very little, if any, similarity exists between judgemental and statistical DIF procedures. Using a judgemental method to establish DIF in test items is no guarantee that the item will be flagged by any statistical DIF procedure. Both methods are thus not fool proof methods of defining the true state of bias in real test data. Power study results from such procedures cannot be accepted as valid.

This study takes the form of a Monte Carlo study in which the experimental conditions are simulated. This form also allows for replications to be made.

In general, the levels of the sample size, test length, item discrimination and item difficulty are crossed to form the experimental conditions within which the performance of the DIF procedures are evaluated. Within each experimental condition, the test item parameters will be determined as spelled out in the item characteristic section below. Examinee samples are randomly generated with abilities from a N(0,1) distribution. Item response strings are simulated. These response strings form the raw data used in the computation of the indices for the different DIF procedures. Within each experimental condition, 100 replications are performed. The values of the index obtained for the different

DIF procedures are then analyzed to determine the effect, if any, of the different experimental conditions.

Assumptions

In this study, certain assumptions are made. The DIF detection procedures used in the study are all designed to address situations where the test measures one underlying trait. To this end, it is assumed in this study that each of the tests used is unidimensional. There are also procedures (e.g. the Mantel-Haenszel and the Logistic Regression procedures) in which total test score is used as a measure of ability. It is therefore assumed that total test score is an efficient measure of the single underlying ability measured by the test. Another assumption made in this study has to do with the number of test items necessary and sufficient for the production of reliable ability measures. It is assumed here that tests of 42 and 66 items will result in reliable ability measures. Ability measures are basically on continuous scales. For the particular abilities measured by each item in the tests, it is assumed that dichotomous (0,1) scoring of items will result in no loss of information. The tests used in this study are assumed to be power rather than speed tests. Thus, each examinee is allowed enough time to answer all the items in each test. Guessing is assumed to be minimal and at the same level for groups of examinees used in the study. Finally, it is assumed that local independence

holds among items in each test - i.e. the performance on one item in no way influences the performance on another item in the test.

Characteristics of Test Items

A. Null (No DIF) study

The average length for subtests of most standardized achievement tests is around 40 items. Thus a 42 item test and a 66 item test could not be considered far removed from reality; and yet they have a difference in test length that is large enough to allow a test length effect to be observed. In the study, the two test lengths (42 items and 66 items) were used. Each of the two tests used in the study is a unidimensional test with uniformly distributed item difficulty (b) parameters. The item difficulty parameters ranged in value from -2.0 to 2.0 with a mean of 0.0. This arrangement is expected to result in normally distributed total test scores for examinee samples with normal (0,1) ability distribution (Anastasi, 1976, page 201; Lord & Novick, 1968, page 392; Nunnally, 1967, page 250).

Theoretically, both the item discrimination, a, and the item difficulty, b, parameters are defined on a scale ranging from $-\infty$ to $+\infty$. However, in real life, negatively discriminating item are not included in any test, and it is not usual to find a test item with a discrimination value greater than 2.0. So that the range of the 'a' parameter found in real life testing situations is from a=0.0 to a=2.0. (see Hambleton & Swaminathan, 1985; page 36). Based on

this range of the discrimination, a, value, a=0.67 to a=1.34 was classified as the range with medium discrimination. Items with discrimination values within this medium range are expected to have ICC's that are neither very "steep" nor very "gradual". Thus, the probability of success, on such items, do not differ drastically for a small change in ability level. Most test items in real life are found to have discrimination values within this medium range. Two discrimination, a, values (a=0.80 and a=1.20) were selected to be used in this study. These two a values were chosen because they represent the "lower" and "upper" extremes of the range of a values within which most test items fall. To examine the effect of these "low" and "high" "a" values, four arrangements, of the two values, were used: In the first arrangement, all the items in the test are given a discrimination value of a=0.80. In the second arrangement, the first half of the test has a=0.80 while the second half of the test has a=1.20. For the third arrangement, the first half-test has a=1.20 while the second half-test has a=0.80. In the fourth arrangement, all the items have a=1.20.

In practice, the difficulty, b, values, found in tests, range from -2.0 to +2.0 (Hambleton & Swaminathan, 1985; page 36). In this study, this difficulty range was equally divided into low, medium and high difficulty regions. In the study the performance of the eight DIF indices at each of the three difficulty regions would be examined. The 42 item test was divided into a first half (items 1 to 21) and a second half (items 22 to 42). In the first half, the first item was given a difficulty, b-value, of -2.0. The b-value of

the other items were obtained by adding 0.20 to the b-value of the item that preceded each. In the second half of the test, item 22 was given a b-value of -2.0. The b-values of other items were obtained in a similar manner to those in the first half test. In a similar way, the 66 item test was also divided into two half tests with the first item in each half test receiving a b-value of -2.0. In this case, the b-values of other items were obtained by adding 0.125 to the b-value of the preceding item.

The pseudo-guessing (c) parameter for all items in each test was fixed at 0.15. The decision to use this fixed value for the c-parameter, for each item, was arrived at because of the problem with convergence when the LOGIST computer program is used to estimate both person and item parameters. Using a fixed c-value improves the chances for convergence on the other parameters when a three-parameter IRT model is used. Also, the researcher intended to keep guessing at a low level for each study group.

The tables of item parameters for the two tests are given in Appendix 1A and 1B.

B. Known DIF study

Two test lengths, 42 items and 66 items, were used in the known DIF study. The item difficulty parameters for each test were assigned in the same way as those in the null study. The only differences were in the difficulty parameters for the items that were given uniform bias. The item discrimination parameters were also arranged in the same way as in the null study except for the

items given nonuniform bias. The pseudo-guessing parameter for all items was fixed at 0.15 as in the null study. Sample sizes and their ability distributions were the same as those used in the null study.

In this study, the b-values in each half test were classified into low, medium and high difficulty levels. Items in the low difficulty level were those with b-values from -2.0 to less than -0.6. Items in the medium difficulty level were those with b-values from -0.6 to 0.6. Items with high difficulty were those with b-values greater than 0.6 to 2.0. This translates into an equal number of items (seven for the half test of the 42 item test and 11 for the half test of the 66 item test) in each of the three levels.

In each test, six items (14% of the 42 item test and 9% of the 66 item test) were biased by altering either their b-values or their a-values for the minority group test. Three items (one in each of the three levels of difficulty defined above) in the first half of the test were given uniform bias by altering the b-value for the minority group. Three items in the second half of the test, with difficulty values corresponding to those items given uniform bias in the first half test, were given nonuniform bias by altering the discrimination, a-value, for the minority group test. For both uniform and nonuniform bias and in each of the biased items, the measure of bias was fixed at an area-between value of 0.6 which is supposed to be a medium bias measure (Rogers & Swaminathan, 1990). To obtain this (0.6) measure of bias, between the ICCs, Raju's

(1988) formulae for altering the a-value and the b-value were used. The item parameter tables are given in Appendix 1C and 1D.

Characteristics of Samples

Because the study is interested in examining the effect of sample size on each of the 8(eight) DIF procedures, three different levels of sample size arrangement were simulated. In each arrangement, the majority to minority sample ratio of 8:3 was maintained. This ratio is in agreement with most real life situations where the size of the focal group is usually smaller than the size of the reference group. The 8:3 ratio is not significantly removed from the 3:1 majority to minority sample ratio found in a number of DIF studies using real data (e.g. Hambleton & Rogers, 1988; Shepard et al., 1985). A 2 to 1 ratio was maintained between samples in the first arrangement (1600/600) and samples in the second arrangement (800/300) and between the second arrangement (800/300) and the third arrangement (400/150). Each majority and each minority sample was generated from N(0,1) ability distribution.

Data Simulation Model and Program

All data were generated based on the three-parameter IRT model. This model represents real life situations, where tests with multiple-choice items are used, better than the one- or two-

parameter models (Traub, 1983). DATAGEN (Carlson, 1983) program was used in all cases. This program is designed for use in generating data for a one-, two- or three-parameter IRT model. Provision is also made, in the program, for the user to input values for the discrimination (a) and the pseudo-guessing (c) parameters, for each simulated item, when the three-parameter model is used to generate data. The program was modified by L. Fleming (a computer programmer formerly with the Computer Services Department, University of Ottawa) to allow for inputting the difficulty (b) value for each of the simulated items and also to hold the 'a', 'b' and 'c' values constant over the 100 replications performed in each experimental condition. The discrimination (a) and difficulty (b) values used, for each of the two tests, were as specified in the "Characteristics of Items" sections above. DATAGEN also makes provision for sample abilities (thetas) to be generated randomly from a normal (0,1) distribution.

Under normal conditions, some guessing is always present in situations were power tests with multiple-choice items are used. In classical test theory, a common index, used to determine the amount of guessing when a multiple-choice item is used, is defined by 1/x (where x is the number of answer options the item has). Thus, in the case of items with four options, the value of the guessing index will be ¼ or 0.25. This definition of the guessing index suggests that all the options were equally attractive or equally repulsive for the group of examinees. This situation, although possible, is a very rare one to observe in real life. In the

extreme case where no examinee knows the correct response to an item, however equally plausible all the options might seem, some of the options become more attractive, to the examinees, than others. Thus, the use of 1/x for the guessing index of an item cannot be defended for the majority of cases in real life. There is hardly any convincing argument for the defense of using any particular arbitrary value for the guessing parameter (c). So that, the only option here is to use a computer program like LOGIST (Wingersky, Barton & Lord, 1982) to estimate the pseudo-guessing (c) parameter for each item. However, it has been found that the "c" parameter is not easy to estimate even when the sample size is large (Lord, 1980). For this reason, in most simulated DIF studies, based on the three-parameter IRT model, a fixed "c" value (which appears reasonable to the researcher) is used for all the items in the test (e.g. Rogers & Swaminathan, 1990, 1989; Merz & Grossen, 1979). Values found in the literature for the fixed "c" are commonly in the range .05 and .25. In this study a fixed pseudo-guessing (c) value of 0.15 is used for each item, in the two test, within each experimental condition.

The ability generation processes for the majority and the minority samples were however started with different random numbers. Also, different random numbers were used as starting values for the ability estimations in each of the 100 replications.

After the ability (theta) value had been simulated for an examinee, for any given item, i, the item parameters for item i and the theta value for the examinee were used in the three-parameter

IRT model to estimate the examinees probability of success on item i. A random uniform deviate between zero and one was generated for the particular examinee; if this number was greater than or equal to the examinee's probability of success, he was given a score of 1 (for correct) on item i; otherwise, he was given a score of 0 (for incorrect) on item i. The (0,1) test data generated from each simulation constituted a set of raw data that was analysed using each of the eight DIF procedures.

The following software were used to estimate indices: Mantel-Haenszel indices were estimated using a program by Ackerman (1986). A Logistic Regression program written by Spray (1991) was used. This program is a two-in-one program which treats uniform DIF and nonuniform DIF in separate analysis. The program therefore uses two separate 1(one) degree of freedom Chi-Square indices. Software to estimate Lord's Chi-Square, Full Chi-Square, Transformed Item Difficulty, SOS2 and SOS4 were written locally.

The Cutoffs

The cutoff values were obtained only in the null (no DIF) study. For the obtained index value for each DIF procedure within each experimental condition, percentiles of .99, .95 and .90 (false-positive rates of .01, .05 and .10) for each item over the 100 replications were computed. THe method of using the distribution of each item, over the 100 replication was adopted rather than that of using the distribution of each replication over

all the items in the test. In the former case, 100 values will be used to define the distribution while in the latter only 42 or 66 cases (depending on the test length) will be used. The value of the index at each of these rates was recorded for each item. For signed indices (e.g. the SOS4 and MH Delta), the absolute value of the indices were used to transform each item's distribution so that a one-tail distribution (from which the percentile values were determined is obtained).

DATA ANALYSIS

A. Null (No DIF) Study:

In this study, three levels of sample size, two levels of test length and four levels of item discrimination arrangement were used as independent variables. When the levels of these independent variable were crossed, 24 experimental conditions were obtained.

To examine the distribution of each index across the 24 experimental conditions, Multivariate General Linear model (Multivariate GLM) and Univariate General Linear model (Univariate GLM) methods in SAS were used to test for differences in the mean values of the percentiles obtained for each item in each test over the 24 experimental conditions. The decision to use the GLM methods stems from the fact that the two test lengths were different; so that half of the cells will have 42 observations while the other half will have 66 observations. Thus MANOVA and ANOVA which are sensitive to differences in cell sizes could not be used. The GLM

procedures are modifications of the MANOVA and ANOVA procedures designed to handle cases where cell sizes are different. The multivariate procedures were mainly undertaken as omnibus test to control for the Type I error.

B. Known DIF Study:

For each index, the three percentile values obtained in the null study, for a particular item, were used individually in comparison with the value of the index obtained in each of the 100 replications, in the known DIF study, for the item that corresponds to the particular null study item. Whenever a replication value was greater than or equal to a percentile value, it was counted as one incident, out of the 100 replication, that the item is flagged as displaying DIF at that particular percentile. The proportion of times (out of the 100 replications) that the item is flagged was subsequently determined.

For procedures with known tests of significance (e.g. procedures with indices distributed as Chi-square), their tabled values were also used on each of the known DIF study results across the 100 replications to determine the proportion of times (out of 100) that the item was flagged.

CHAPTER V

Results and Discussion

Results of the study are presented and discussed in this chapter. The DIF procedures are treated individually. However, because of the similarity of the results of the SOS2 and the SOS4, their results are presented as one. For each procedure, the results obtained for the null study are presented first followed by the results for the power study. The null study results are presented for the distribution of the index in terms of (i) the obtained means, in each experimental condition, for the three percentile (90^{th}, 95^{th} and 99^{th}) values; and (ii) the GLM results on the effects of the independent variables on the means of the percentile values. For the power study, the results for each DIF procedure will be presented in terms of (i) the number of times, out of 100 replications (given in proportion form) that an item, simulated with bias (uniform and nonuniform) was flagged by an index; and (ii) the number of false-positive identification (given in proportion over 100 replications of the simulated unbiased items) for each index. For procedures with known tests of significance, results (at alpha = .05) for correct and false-positive identifications, will also be presented and compared with the results obtained when the 95^{th} percentile cutoff from the null study. Both the null study and the power study results are discussed in terms of the nature of the particular procedure and also in terms of results from past studies.

For each item, three percentiles were estimated, over the 100 replications, for each index. Thus for the 42 item test, there were 12 experimental conditions (four discrimination levels by three sample sizes). In each condition, there were 42 estimates per percentile per index. Similarly, with the 66 item tests, there were 66 estimates, per percentile, per index, in the 12 experimental conditions. Means for the 42 and 66 values (as the case may be), per percentile, for each index, are reported for the null study. Multivariate GLM analysis on these means, using sample size, test length, and discrimination arrangement as independent variables, were performed. After a significant multivariate effect, all necessary univariate GLM analyses were performed by index and percentile. A .01 alpha level was used in all multivariate and univariate GLM analyses. Whenever post hoc analysis was necessary, the Scheffe procedure (with the alpha level adjusted to prevent an inflation of the Type I error rate) was used.

For ease of presentation and discussion of results, the following notation system is adopted for the different levels of the independent variables: Sample size, test length, discrimination (a-values) and difficulty (b-values). For sample size, S1, S2 and S3 represent the sample size (majority/minority) arrangements, 1600/600, 800/300 and 400/150 respectively. For test length, TL1 and TL2 represent the 66 item test and the 42 item test respectively. For discrimination (a-value) arrangements, D1, D2, D3, and D4 represent the discrimination (a-value) arrangements: a) Where all items in the test had a-values of 0.8, b) Where all items

in the first half of the test had a-values of 0.8 and all items in the second half of the test had a-values of 1.2, c) Where all items in the first half of the test had a-values of 1.2 and all items in the second half of the test had a-values of 0.8, and d) Where all items in the test had a-values of 1.2 respectively. For the item difficulty (b-value) arrangements (i.e. for the power study), Low Diff, Medium Diff and High Diff represent a) the item difficulty (b-values) ranging from -2.00 to -0.66, b) where all the items had difficulty (b-values) ranging from -0.65 to 0.66, and c) where all the items had difficulty (b-values) ranging from 0.67 to 2.00 respectively. The results for procedures based on the IRT model are presented first. This is followed by the results of the non-IRT procedures.

IRT Procedures

The Sum of Square (SOS) Indices

As was mentioned earlier, the results obtained for the SOS2 and the SOS4 were very similar. Thus, only one (that for the SOS2) is presented. After simulating the (0,1) test data for the different experimental conditions in the null study, a problem was encountered when the (0,1) test data were subjected to LOGIST analysis for the estimation of parameters and probabilities of success. It was observed that for the two smaller sample size arrangements (S2 and S3), LOGIST failed to converge, and so the program halted, in the majority of cases. On the other hand, stable

estimates were obtained only for the 1600/600 sample in each of the experimental conditions. Because of the unavailability of other IRT estimation programs, like BILOG, a decision was taken to drop the two smaller samples (S2 and S3) from the study of all the IRT based DIF procedures. So only one sample size arrangement (S1) was used for the SOS indices.

Distribution study results (SOS2)

The GLM analysis (see Table 1) produced significant test length and discrimination effects for P_{90}, P_{95} and P_{99}. The results for P_{90} and P_{95} were identical: For each of these two percentile values, TL1 produced means that were significantly larger than the means produced by TL2. The discrimination effects on P_{90} and P_{95} indicated that D1 produced means that were significantly larger than the means for any of the other discrimination arrangements (D2, D3 or D4). Also D2 and D3 each produced means that were significantly larger than the means for D4. The test length result for P_{99} was similar to those for P_{90} and P_{95}: TL1 produced significantly larger means than those produced by TL2. For the discrimination effect on P_{99}, only D2 was found to produce significantly larger means than D4.

Power study results (SOS2)

In general, the bias detection rate of the index was relatively low. For uniform bias, the rate was .32 while the rate

Table 1

Means by Percentile and Tests of Significance for Different Independent Variables for SOS2

	DISCRIMINATION											
	D1			D2			D3			D4		
Test Length	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}
TL1	5.636	6.979	9.271	4.412	6.055	11.130	4.409	6.096	8.921	3.227	3.702	6.976
TL2	5.269	6.728	8.082	4.270	5.900	7.399	4.283	5.766	7.816	3.150	3.657	5.507

Significance found for test length and Discrimination for P_{90} P_{95} and P_{99}

P_{90} & P_{95} Test Length Long > Short
Discrimination
D1 > D2, D3, D4 and
D2, D3 > D4

P_{99} Test Length Long > Short
Discrimination
D2 > D4

for nonuniform bias was .01. The overall false-positive rate was .053. See Tables 2 and 3.

The bias detection rates for TL1 and TL2 were slightly different when the items were uniformly biased: TL1 produced a detection rate of .37 which was slightly larger then .34, the detection rate for TL2. When the bias was nonuniform both test lengths (TL1 and TL2) produced rates lower than .01. The false-positive rates for TL1 and TL2 were similar at .053.

Considering only the case of low discrimination (a = 0.8) and high discrimination (a = 1.2), that is the case of D1 compared to D4, with uniform bias D4 produced a bias detection rate of .49 while D1 had a rate of .36. When the bias was nonuniform both D4 and D1 had identical bias detection rates of .01. The false-positive rates for D4 and D1 were .049 and .064 respectively.

With uniformly biased items, items with medium difficulty had the highest detection rate of .53. Items with low difficulty had the second highest bias detection rate of .49. Items with high difficulty were very poorly detected at .04. When the bias was nonuniform, the SOS2 was unable to detect items with medium difficulty (detection rate was .00). The bias detection rate for nonuniform bias was identical for items with low and high difficulties (rate of .01 in each case). Medium difficulty items had the highest false-positive rate of .091. Low and high difficulty items had false-positive rates of .036 and .030 respectively.

Table 2

Different Detection Rates (in Proportions) by Levels of Independent Variables for SOS2 (AT P_{95} Cutoff Value)

		DISCRIMINATION								
		D1		D2		D3		D4		
Kind of Bias	Difficulty Level	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total
UNIFORM	Low	.50	.45	.70	.58	.05	.11	.77	.76	.49
	Medium	.64	.48	.85	.75	.00	.11	.79	.63	.53
	High	.02	.05	.06	.13	.00	.00	.00	.01	.04
	Total	.39	.33	.54	.49	.02	.07	.52	.06	.32
NONUNIFORM	Low	.00	.01	.00	.00	.01	.01	.00	.02	.01
	Medium	.00	.00	.00	.00	.00	.00	.00	.00	.00
	High	.00	.01	.00	.00	.01	.00	.00	.02	.01
	Total	.00	.01	.00	.00	.01	.00	.00	.01	.01

Table 3

False-Positive Rates (in Proportions) by Levels of Independent Variables for SOS2 (AT P_{95} Cutoff Value)

		DISCRIMINATION								
		D1		D2		D3		D4		
Half of Test	Difficulty Level	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total
	Low	.018	.058	.140	.025	.001	.006	.011	.002	.032
First	Medium	.108	.095	.161	.152	.005	.015	.124	.107	.096
	High	.023	.035	.059	.053	.006	.003	.023	.020	.028
	Total	.050	.063	.120	.077	.004	.006	.056	.043	.052
	Low	.021	.183	.000	.000	.053	.047	.009	.003	.040
Last	Medium	.090	.073	.003	.002	.157	.157	.102	.102	.086
	High	.023	.028	.011	.007	.059	.062	.033	.032	.032
	Total	.015	.095	.005	.003	.090	.089	.048	.046	.053
	Grand Total	.048	.079	.063	.040	.047	.048	.052	.045	.053

The results in Table 1 clearly indicate that test-length consistently affected the size of each of the three percentiles of the SOS2 index. A positive relationship was observed between test length and the size of the percentiles; with longer tests consistently producing significantly larger values, of each of the three percentiles, than shorter tests. From the definition of the SOS2 index, disregarding sample size, two quantities determine the size of the index. The first is the sum of the squared differences in the P(θ) values. This is given by:

$$\sum_{i=1}^{n_R+n_F} [P_{iR}(\theta) - P_{iF}(\theta)]^2$$

and the other is the variance of these values:

given by:

$$\sigma^2_{P_{iR}(\theta)} - P_{iF}(\theta).$$

If the former increases (everything else being equal), the index will increase, and if it decreases, the index will decrease. on the other hand, if the latter (the variance) increase (everything else being equal) the index will decrease and if it decreases the index will increase. Since the obtained test length effects indicates an increase in the value of the percentiles with increase in test length, one or both of the following conditions should be responsible: As test length increase, either the sum of the squared

differences in P(θ) values increases or the value of the variance of the P(θ) differences decreases.

When a test is longer, theta estimates would tend to be more accurate tending to reduce the numerator of the SOS2. It should also reduce considerably, the variance of these differences thus tending to make the SOS2 have a higher measure.

The means of the a-values in each of D1, D2, D3 and D4 were 0.8, 1.0, 1.0, and 1.2 respectively. From the obtained results (presented in Table 1) it appears that (i) when the mean discrimination values of two tests are different, the test with the smaller mean a-value produces significantly larger P_{90} and P_{95} values than the test with the larger mean a-value, and (ii) when the mean a-value of two tests are the same, no significant differences are observed in the P_{90} or P_{95} values of the two tests.

In general, the higher the a-value of an item, the steeper the item characteristic curve (ICC) of that item. Thus, for any group of examinees, if the item's a-value is high, most of the probabilities of success, P(θ), associated with the ability levels of the examinees, will be concentrated closer to the lower asymptote (in this case, the pseudo-guessing parameter) for less abled examinees, and closer to 1.0 (the upper asymptote) for the more abled examinees. For each of the two (reference or focal) groups, given an item with a high a-value, the P(θ)'s for each group will cluster around the upper and lower asymptotes. in such a situation, the difference between the P(θ)'s, for the two groups, at any given ability level, will be very close to, if not exactly

equal to zero. Thus even when such differences are squared and summed over all the ability levels in the two groups, the sum is expected to be small. This implies that the SOS2 index for such an item (i.e. with a high a-value) will be small. On the other hand, when the a-value is low, the ICC will have a fairly gentle slope for each group. Here, the chances are greater that the group's $P(\theta)$ differences will be further away from zero. This will produce a larger sum of squared $P(\theta)$ differences when taken over the range of abilities in both groups. Thus with a small a-value, the SOS2 index for such an item is expected to be large.

There are, so far, very few comparative studies in the literature that have included the SOS2 indices. Two such studies are Shepard, Camilli, and Williams (1985) and Skagg and Lissitz (1988). The study by Shepard, Camilli, and Williams (1985) was in two parts. In the first study, the SOS2 and SOS4 were used in a crossvalidation study to establish the bias criterion in the test data set. The established criterion was then used to test the power of bias detection of other DIF Procedures. In the second study, which was a simulation study, both the SOS2 and SOS4, together with other DIF procedures, were examined for their power to identify known uniform bias built into the simulated test data. in this second study, the SOS2 and SOS4 flagged four and five items respectively out of the 18 biased items. This gives 9 bias detection rate of .22 and .28 respectively for the SOS2 and SOS4. However, when only moderately biased items were used, the bias detection rates of the SOS2 and SOS4 changed to .33 and .44

respectively. In the current study, only moderately biased items were used. The bias detection rate for uniformly biased items in the current study was .32 for both the SOS2 and SOS4. The false-positive rates of the SOS indices were very low in both studies. Clearly this result is similar to that obtained for the SOS2 (with only moderate uniformly biased items) in the Shepard, Camilli, and Williams (1985) study.

Skagg and Lissitz (1988), were mainly concerned with the reliability and validity of the different DIF procedures. There is, therefore, not much comparison that could be made with the current study. However, one finding of Skagg and Lissitz (1988), that is of major importance to the current study, is that they found sample size to affect the size of the SOS2 and SOS4 indices; with larger samples producing larger values of the indices. The current study was designed to include sample size as an independent variable. For reasons given earlier, the sample size variable was excluded from the study of the SOS indices.

The Lord's Chi-Square Index

For the same reasons given for the SOS indices, the two smaller sample size arrangements (S2 and S3) were excluded from the study of Lord's chi-square.

Distribution study results (Lord's χ^2

With respect to the two independent variables (test length and discrimination) in the study of Lord's chi-square, no significant

results were found for any of the three percentiles (P_{30}, P_{35} and P_{40}) involved (See Table 4).

Power study results (Lord's χ^2)

For both uniform and nonuniform bias Lord's chi-square produced a perfect bias detection rate of 1.00. In general, the false-positive rate of .081 was higher than any of those for the SOS2 and SOS4, the other IRT based DIF indices in the study. These results are presented in Table 5.

The false-positive rate for Lord's chi-square was negatively affected by test length: The longer test produced a lower false-positive rate than the shorter test. The false-positive rate for the longer test was .071 while that for the shorter test was .092. A slight negative relationship was also observed between the false-positive rate and item discrimination (a-value): Items with low discrimination (i.e. items in D1) produced a false-positive rate of .086 while items with high discrimination (i.e. items in D4) produced a false-positive rate of .082. Item difficulty also displayed a negative relationship with false-positive rate: For low, medium and high difficulty items, the false-positive rates were .088, .085 and .071 respectively (See Table 6).

Although sample size was excluded from the study of Lord's chi-square, it is expected that the index, being a chi-square index, will be affected by sample size. However, it is worth noting that the performance of the index, in this study, was outstanding. These results were not surprising since the index involves the

Table 4

Means by Percentile and Tests of Significance for Different Independent Variables for LORD'S CHI SQUARE

	DISCRIMINATION											
	D1			D2			D3			D4		
Test Length	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}
TL1	3.000	4.023	6.131	3.274	4.285	6.756	3.074	4.160	6.402	3.160	4.052	5.777
TL2	3.536	4.864	6.357	3.178	4.227	5.665	3.039	4.010	5.668	2.965	3.968	5.529

No significance found!

Table 5

Different Detection Rates (in Proportions) by Levels of Independent Variables for LORD'S CHI SQUARE (AT P_{95} Cutoff Value)

		DISCRIMINATION								
		D1		D2		D3		D4		
Kind of Bias	Difficulty Level	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total
UNIFORM	Low	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Medium	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	High	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Total	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NONUNIFORM	Low	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Medium	1.00	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	High	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Total	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 6

False-Positive Rates (in Proportions) by Levels of Independent Variables for LORD'S CHI SQUARE (AT P_{95} Cutoff Value)

		DISCRIMINATION								
		D1		D2		D3		D4		
Half of Test	Difficulty Level	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total
	Low	.063	.220	.057	.077	.055	.102	.094	.097	.096
First	Medium	.073	.083	.067	.097	.067	.095	.082	.090	.082
	High	.078	.072	.060	.068	.059	.078	.059	.073	.068
	Total	.091	.125	.061	.081	.060	.092	.078	.087	.082
	Low	.073	.068	.060	.077	.067	.127	.088	.075	.079
Last	Medium	.086	.087	.073	.073	.072	.123	.086	.092	.087
	High	.071	.058	.069	.062	.070	.118	.062	.078	.074
	Total	.077	.071	.067	.071	.070	.123	.079	.082	.080
	Grand Total	.074	.098	.064	.076	.065	.108	.079	.085	.081

simultaneous comparisons of the difficulty (b) and the discrimination (a) parameters, of the studied item, for the two groups.

Shepard, Camilli, and Williams (1985) found Lord's chi-square not effective in detecting weakly biased items and somewhat effective in detecting moderately biased items. The bias detection rates for weak and moderate biases were .11 and .56 respectively. The false-positive rate for Lord's chi-square reported in Shepard, Camilli and Williams (1985) was .028. Although the current study involved only moderate bias, the result for moderate bias in Shepard, Camilli, and Williams (1985) is significantly different from the results obtained in this study for Lord's chi-square. Whereas the bias detection rate for Lord's chi-square was 1.00 in the current study, the Shepard, Camilli and Williams (1985) result that corresponds to the current study result was a detection rate of .56. Also the two studies differ in the false-positive rates reported for Lord's chi-square. The current study produced a false-positive rate of .081 as against .028, the rate observed by Shepard, Camilli and Williams (1985). In the current study, after the test data were generated for the specific item and person parameters, for the groups, LOGIST was used in the reparametization. Here, the problem of unstable parameter estimates was observed for S2 and S3 even for the 66 item test. Clearly the condition involving S2 (the sample with 800 and 300 majority and minority samples respectively) and the 66 item test is not far removed from the Shepard, Camilli and Williams (1985) condition of

1,000 and 300 majority and minority samples and 54 items. Thus it seems likely that they were also troubled by sample size. Such a situation would definitely strongly affect the Lord's chi-square results.

Skagg and Lissitz (1988) found that sample size had a very strong effect on the Lord's chi-square index. Larger samples consistently produced larger values of the index. In a baseline study, they used samples that were about 3 times larger than the samples used in two of the conditions they studied. This resulted in baseline values that were larger than the index values produced by any of the two smaller samples. As was mentioned earlier, this was not surprising since the index is distributed as a chi-square which is known to be sensitive to sample size.

Non-IRT Procedures

The Transformed Item Difficulty (TID) Index

The TID is one DIF index that has received mixed reviews based on results from comparative studies. Seong and Subkoviak (1987) found the TID performed worse than either the IRT-Z index or the full chi-square index. Shepard, Camilli, and Williams (1985) found it inadequate. Rudner, Getson, and Knight (1980) found the TID performed well on items with uniform bias but very poorly on items with nonuniform bias. Merz, and Grossen (1979) found the index gave a better performance than any of the other indices they considered.

Distribution study results (TID)

Results from this study indicated that the three percentiles (P_{33}, P_{35}, and P_{99}) of the TID index were significantly affected by sample size and test length. A significant interaction effect, between sample size and test length was also observed for P_{99} (See Table 7).

Following the significant sample size and test length interaction for P_{99}, the simple effect analyses produced the following results: For the longer test, the P_{99} means for S3 and S2 were not significantly different. However, each of S3 and S2 produced P_{99} means that were significantly larger than those for S1. For the shorter test, the P_{99} means for S3 were significantly larger than the means for either S2 or S1. Also the means produced by S2 were significantly larger than the means for S1. At each of the three sample size (S1, S2, and S3) levels, TL1 produced P_{99} means that were significantly larger than those produced by TL2.

Power study results (TID)

The performance of the TID index was significantly better on items with uniform bias than on items with nonuniform bias (see Table 8). Overall, the bias detection rates for uniform and nonuniform bias were .62 and .05 respectively. The false-positive rate (see Table 9) was .072.

Sample size displayed a positive relationship with uniform bias detection rate for the TID. When the items were uniformly biased, the detection rates were .92, .64, and .29 for S1, S2 and

Table 7

Means by Percentile and Tests of Significance for Different Independent Variables for TID

		DISCRIMINATION											
		D 1			D 2			D 3			D 4		
Test Length	Sample Size	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}
	S1	1.688	2.376	4.280	1.678	2.549	4.960	1.755	2.577	5.130	1.643	2.559	5.161
TL1	S2	2.578	3.757	7.230	2.688	4.058	9.140	2.517	3.654	8.180	2.486	3.801	8.560
	S3	4.575	6.717	15.570	4.323	6.533	16.210	4.673	6.891	16.110	4.862	7.483	26.660
	S1	1.680	2.351	3.450	1.682	2.427	3.690	1.539	2.430	3.880	1.556	2.387	3.940
TL2	S2	2.497	3.710	5.480	2.392	3.599	5.520	2.389	3.693	5.690	2.558	3.964	6.360
	S3	3.796	5.513	9.750	4.155	6.493	11.310	4.163	6.498	10.690	4.231	6.475	11.360

Significant found for both sample size and test length for P_{90} P_{95} P_{99}

Significant found for test-length sample size interaction for P_{99}

P_{90}
Large < Medium < Small
Long test > Short

P_{95}
Large < Medium < Small
Long test > Short

P_{99}
Simple Effects - After Interaction
Long Test: Large Sample Size < Medium & Small Sample Sizes
Short Test: Large Sample Size < Medium < Small
Large Sample Size: Long Test > Short Test
Medium Sample Size: Long Test > Short Test
Small Sample Size: Long Test > Short Test

Table 8

Different Detection Rates (in Proportion) by Levels of Independent Variables for TID (at P_{95} Cut-off Value)

		UNIFORM									NONUNIFORM								
		DISCRIMINATION									DISCRIMINATION								
		D1		D2		D3		D4			D1		D2		D3		D4		
Difficulty Level	Sample Size	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total
Low	S1	.94	.87	.84	.80	.96	.93	.94	.96	.92	.05	.05	.01	.01	.04	.05	.02	.00	.03
	S2	.62	.53	.45	.46	.77	.77	.63	.70	.62	.05	.10	.01	.00	.03	.03	.01	.00	.03
	S3	.25	.30	.25	.16	.40	.37	.29	.32	.29	.04	.08	.02	.05	.05	.06	.03	.06	.05
	Total	.60	.57	.51	.47	.71	.69	.62	.66	.61	.05	.08	.01	.02	.04	.05	.02	.02	.04
Medium	S1	.95	.92	.84	.86	.90	.89	.90	.96	.90	.05	.02	.00	.00	.04	.06	.00	.00	.02
	S2	.70	.61	.55	.58	.67	.64	.53	.53	.60	.04	.04	.00	.01	.06	.03	.01	.00	.02
	S3	.24	.35	.22	.16	.23	.28	.13	.24	.23	.04	.01	.02	.01	.01	.03	.00	.00	.02
	Total	.63	.63	.54	.53	.60	.60	.52	.58	.58	.04	.02	.01	.01	.04	.04	.00	.00	.02
High	S1	.95	.99	.95	.98	.93	.90	.86	.92	.93	.01	.23	.04	.03	.06	.15	.02	.06	.08
	S2	.77	.83	.70	.82	.62	.73	.57	.60	.71	.05	.09	.04	.06	.13	.11	.03	.02	.07
	S3	.39	.48	.39	.34	.28	.36	.32	.34	.36	.13	.06	.11	.06	.14	.08	.08	.08	.09
	Total	.70	.77	.68	.71	.61	.66	.58	.62	.67	.06	.13	.06	.05	.11	.08	.04	.05	.08
	Grand Total	.64	.66	.58	.57	.64	.65	.57	.62	.62	.05	.08	.03	.03	.05	.06	.02	.02	.05

Table 9

False-Positive Rates (in Proportion) by Levels of Independent Variables for TID (at P_{95} Cutoff Value)

Difficulty Level	Discrimination	SAMPLE SIZE S1 TL1 First	S1 TL1 Last	S1 TL2 First	S1 TL2 Last	S2 TL1 First	S2 TL1 Last	S2 TL2 First	S2 TL2 Last	S3 TL1 First	S3 TL1 Last	S3 TL2 First	S3 TL2 Last	Total
Low	D1	.047	.057	.088	.057	.038	.041	.053	.053	.058	.055	.065	.077	.057
	D2	.034	.035	.057	.040	.035	.034	.058	.042	.095	.090	.090	.072	.049
	D3	.077	.081	.070	.078	.030	.034	.055	.042	.043	.051	.087	.087	.061
	D4	.021	.024	.037	.047	.028	.050	.051	.035	.091	.096	.087	.075	.053
	Total	.049	.049	.063	.056	.033	.040	.054	.043	.072	.073	.082	.078	.055
Medium	D1	.004	.004	.010	.008	.000	.000	.003	.003	.002	.000	.007	.007	.004
	D2	.003	.000	.002	.000	.002	.000	.003	.000	.009	.004	.007	.008	.003
	D3	.004	.006	.010	.008	.000	.001	.000	.000	.000	.002	.015	.008	.005
	D4	.000	.000	.000	.000	.003	.000	.002	.003	.006	.006	.010	.008	.002
	Total	.003	.003	.006	.004	.001	.000	.002	.002	.004	.003	.010	.008	.004
High	D1	.157	.167	.151	.158	.151	.156	.125	.133	.120	.143	.165	.167	.149
	D2	.106	.146	.118	.155	.102	.152	.147	.178	.182	.207	.243	.188	.152
	D3	.250	.133	.198	.138	.135	.118	.135	.123	.163	.128	.205	.172	.158
	D4	.134	.145	.172	.165	.168	.157	.127	.145	.184	.196	.200	.202	.166
	Total	.161	.148	.160	.154	.139	.146	.134	.145	.162	.168	.178	.182	.156

S3 respectively. When the items were nonuniformly biased, the detection rates dropped considerably to .04, .04, and .05 for S1, S2, and S3 respectively. The corresponding false-positive rates for S1, S2 and S3 were .072, .062, and .085 respectively.

Test length did not appear to affect the detection rates nor the false-positive rates of TID. For uniform bias detection, the rates were much better than for nonuniform bias.

The uniform bias detection rate was slightly better, at .65, for items with low discrimination (i.e. D1) than for items with higher discrimination (i.e. D4) which had a rate of .60. When the items were nonuniformly biased, the detection rates were .07 and .02 for D1 and D4 respectively. False-positive rates of D1 and D4 were .070 and .074 respectively.

With uniformly biased items, low, medium and high difficulty items had rates of .61, .58, and .67 respectively. For nonuniformly biased items the rates were .04, .02 and .08 respectively for low, medium, and high difficulty items. The corresponding false-positive rates for low, medium and high difficulty items were .055, .004, and .156 respectively.

One of the main criticisms against the TID DIF index is that it confounds DIF with item discrimination. Angoff (1982) suggested the use of the item-test correlation to correct this confounding problem. This study used the Angoff (1982) suggested modification. From the results of the distribution study, and also from the power study, it is clear that there was no confounding of bias with discrimination. In the distribution study, no significance was

found for item discrimination. If discrimination was confounded with bias, the distribution study would have shown an effect of item discrimination on the three percentiles. In the power study also, with the said confounding effect, one would have expected a significant difference in the bias detection rates for low and high discrimination items. This was clearly not the case in this study.

In Shepard, Camilli, and Williams (1985), the results for moderately biased items, in study two, is the only one that could be meaningfully compared with any part of the results obtained in the bias detection power section of the current study. This is so because the study by Shepard, Camilli, and Williams (1985) only dealt with power of detecting uniformly biased items. They also included weakly and moderately biased items in their study. In the current study, however, only items with moderate bias were examined. For the moderately biased items in their study, Shepard, Camilli, and Williams (1985) observed a bias detection rate of .67 for the TID procedure. The false-positive rate was .25. In the current study, for uniformly biased items the TID bias detection rate was .62 with an overall false-positive rate of .072. The major difference in these two results is in terms of the false-positive rates obtained in the two studies. However, when they used a modified TID (in the form of a residualized index), their observed detection rate stayed the same (.67); but the false-positive rate dropped to .111 which is not very much different from the .072 false-positive rate observed in this study.

Rudner, Getson and Knight (1980) found that when items were uniformly biased the TID bias detection results correlated highly (.80 and .87) with the generated bias state of the data. However, when the bias in the data was purely nonuniform (i.e. only in the a-value), the correlation, with the generated bias state, dropped to .52 and .45. These correlations do not say much about the actual detection rates of uniformly and nonuniformly biased items. However, both studies found no significant effects of test length on the bias detection performance of the TID.

In Merz and Grossen (1979) when the overall test difficulty was .80, the bias detection rate of the TID was 1.00 and .92 respectively for 10% and 20% of the items biased in a test of 60 items. The corresponding false-positive rates in the two situations were the same, at .074 in each case. When the overall test difficulty was .60, the bias detection rates of the TID were .83 and .92 respectively for 10% and 20% of the test items biased. The corresponding false-positive rates in the two situations were .074 and .13 respectively. Merz and Grossen used only uniformly biased items. The bias detection rates observed by Merz and Grossen were, in all cases, higher than the rate of .62 observed for uniformly biased items in the present study. This may be explained in terms of the differences in the forms of the TID used in the two studies. However a common pattern appears to have surfaced in the two studies when the percentage of biased items in the test is less than 20%. In the current study, the percentage of biased items was less than 20% in all cases. For low difficulty items, a detection

rate of .61 was observed for uniform bias. When the difficulty was moderate, the detection rate dropped to .58 for uniform bias. In the Merz & Grossen (1979) study, when the difficulty level of the test rose from low (80%) to moderate (60%) the detection rate dropped from 1.00 to .83 for the test with 10% of the items biased. The overall false-positive rates for the two studies were similar.

Seong and Subkoviak (1987) also used a signed form of the TID. The signed form produced a bias detection rate of .80 and a false-positive rate of .025. Although the modified form of the TID used in the current study was the same used by Seong and Subkoviak, the methods by which bias was built into the test data were different. However, the two results do not show much disagreements in terms of bias detection or False-positive rates.

The Full Chi-Square (χ^2_{full})

The mean percentile values obtained for the full chi-square were quite small considering this had a six degree of freedom chi-square distribution. These mean percentile values and the GLM results are presented in Table 10. Significance was found for discrimination for $P_{\theta\theta}$ and $P_{\theta s}$; and for test length and discrimination for P_{ss}.

Considering the two main discrimination levels D1 (items with low discrimination) and D4 (items with high discrimination), the $P_{\theta\theta}$, $P_{\theta s}$, and P_{ss} results indicate that the means for D4 were, in each case, significantly larger than the means for D1. If these findings are replicated in other studies it would mean a serious problem for

Table 10

Means by Percentile and Tests of Significance for Different Independent Variables for FULL CHI-SQUARE

Test Length	Sample Size	DISCRIMINATION D1 P_{90}	D1 P_{95}	D1 P_{99}	D2 P_{90}	D2 P_{95}	D2 P_{99}	D3 P_{90}	D3 P_{95}	D3 P_{99}	D4 P_{90}	D4 P_{95}	D4 P_{99}
TL1	S1	2.687	3.885	5.332	2.882	4.139	5.765	2.660	3.719	5.340	3.054	4.245	5.734
	S2	2.807	3.876	6.342	2.786	4.065	5.398	2.951	4.231	6.048	3.362	4.625	6.368
	S3	S.862	4.012	5.868	2.992	4.245	5.977	2.950	4.401	6.063	3.118	4.278	6.294
TL2	S1	2.670	3.519	5.490	2.920	4.039	6.888	2.785	4.066	6.624	2.988	4.152	6.868
	S2	2.786	3.863	6.336	2.749	3.821	6.117	2.817	3.837	6.237	3.016	4.173	6.922
	S3	2.873	3.929	6.606	2.769	3.833	6.176	2.981	4.152	6.659	3.110	4.343	7.089

Significance found for discrimination for P_{90} — For P_{90}: D4 > D1, D2, D3

Significance found for discrimination for P_{95} — For P_{95}: D4 > D1

Significance found for test length and discrimination for P_{99} — For P_{99}: Long Test > Short Test; D4 > D1

the full chi-square index. Its chi-square distributional claims would have to be re-examined. The results imply that the P_{gs} value would be larger for items with high discrimination values; and lower for items with low discrimination values. This definitely does not translate into the normal situation where the chi-square tabled value at $\alpha = .05$ is used as the cutoff value. The tabled cutoff value depends only on the degrees of freedom which will be the same for both items.

Test length displayed a positive relationship with the means for P_{gg}. This finding may be a result of the intervals for the long test having greater variability than those for the shorter test.

Bias detection power study (χ^2_{full})

In general, the performance of the full chi-square index, in bias detection, was poor. The results are presented in Table 11, for bias detection, and Table 12, for false-positive rates. For items with nonuniform bias, the detection rate was .41 which was better than the detection rate for uniform bias which was .25. This result was a big surprise because this study used the signed form of the full chi-square index which tends to ignore the existence of nonuniform DIF. This surprising result seems to be related to the difficulty levels of the items. For uniform bias, the bias detection rates for low, medium, and high difficulty items were .19, .19, and .38 respectively. For uniform bias therefore the rates only change when the difficulty changes from medium to high. For nonuniform bias, however, the bias detection rates for low,

Table 11

Different Detection Rates (in Proportion) by Levels of Independent Variables for FULL CHI-SQUARE (at P_{95} Cutoff Value)

		UNIFORM									NONUNIFORM								
		DISCRIMINATION									DISCRIMINATION								
		D1		D2		D3		D4			D1		D2		D3		D4		
Difficulty Level	Sample Size	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total
Low	S1	.16	.11	.13	.06	.57	.46	.55	.40	.31	.15	.32	.09	.20	.24	.44	.13	.15	.22
	S2	.11	.05	.03	.05	.34	.23	.35	.20	.17	.09	.16	.06	.08	.09	.17	.06	.07	.10
	S3	.03	.03	.03	.04	.15	.14	.19	.07	.09	.02	.08	.04	.05	.09	.12	.05	.05	.06
	Total	.10	.06	.06	.05	.35	.28	.36	.22	.19	.09	.17	.06	.11	.14	.24	.08	.09	.13
Medium	S1	.19	.18	.14	.15	.29	.30	.39	.31	.24	.62	.77	.68	.65	.69	.85	.70	.82	.72
	S2	.16	.14	.14	.16	.23	.13	.24	.16	.17	.36	.52	.45	.41	.42	.46	.50	.43	.44
	S3	.12	.09	.14	.18	.14	.15	.21	.14	.15	.15	.20	.16	.24	.17	.22	.18	.23	.20
	Total	.16	.14	.14	.16	.22	.19	.28	.20	.19	.38	.50	.43	.43	.43	.50	.46	.49	.45
High	S1	.38	.40	.32	.33	.47	.62	.49	.52	.44	.15	.92	.83	.93	.88	.97	.88	.96	.82
	S2	.32	.33	.36	.31	.30	.51	.36	.35	.36	.55	.78	.68	.64	.66	.72	.61	.69	.66
	S3	.18	.26	.27	.28	.29	.38	.39	.35	.34	.34	.39	.45	.48	.43	.44	.36	.50	.41
	Total	.29	.33	.32	.31	.35	.50	.41	.41	.38	.35	.70	.69	.68	.69	.71	.62	.68	.66
	Grand Total	.18	.18	.17	.17	.21	.22	.35	.28	.25	.27	.46	.39	.41	.42	.48	.39	.42	.41

Table 12

False-Positive Rates (in Proportion) by Levels of Independent Variables for FULL CHI-SQUARE (at P_{95} Cutoff Value)

		SAMPLE SIZE												
		S1				S2				S3				
Difficulty Level	Discrimination	TL1 First	TL1 Last	TL2 First	TL2 Last	TL1 First	TL1 Last	TL2 First	TL2 Last	TL1 First	TL1 Last	TL2 First	TL2 Last	Total
Low	D1	.040	.042	.097	.107	.024	.024	.040	.050	.009	.012	.015	.013	.004
	D2	.038	.027	.082	.055	.041	.031	.043	.050	.011	.014	.025	.023	.037
	D3	.015	.021	.158	.155	.016	.022	.038	.043	.021	.023	.048	.057	.051
	D4	.018	.013	.037	.040	.029	.028	.033	.045	.010	.012	.015	.013	.025
	Total	.028	.026	.094	.089	.028	.026	.039	.047	.013	.015	.026	.027	.038
Medium	D1	.126	.112	.228	.222	.087	.082	.102	.107	.040	.055	.067	.045	.093
	D2	.077	.103	.150	.167	.095	.105	.110	.137	.070	.071	.073	.087	.104
	D3	.100	.085	.288	.250	.069	.070	.123	.118	.060	.058	.123	.105	.121
	D4	.066	.043	.150	.157	.077	.077	.067	.083	.061	.054	.082	.093	.084
	Total	.093	.086	.204	.199	.082	.084	.105	.111	.058	.060	.086	.083	.105
High	D1	.269	.253	.363	.380	.242	.238	.253	.218	.166	.172	.212	.208	.256
	D2	.210	.248	.280	.307	.265	.284	.298	.313	.202	.241	.238	.278	.265
	D3	.256	.214	.403	.365	.263	.223	.298	.270	.244	.223	.300	.285	.280
	D4	.252	.231	.325	.303	.275	.271	.247	.212	.228	.241	.245	.292	.261
	Total	.247	.237	.343	.334	.261	.254	.274	.253	.210	.219	.248	.266	.263
	Grand Total	.123	.116	.214	.207	.124	.121	.139	.137	.094	.098	.120	.125	.135

medium and high difficulty items were .13, .45, and .66 respectively. For nonuniform bias, therefore, the bias detection rate shows a significant positive relationship with the item difficulty levels. What seems to be happening here is that bias tends to be confounded with difficulty level when the item difficulty changes from medium to high for uniformly biased items. When the item is nonuniformly biased, the confounding of item bias with item difficulty is apparent between any two of the levels of difficulty. This is clearly a weakness in the index. The false-positive rates also tend to support the claim of confounding bias with item difficulty. The false-positive rates for low, medium and high difficulty items were .038, .105 and .263 respectively.

The observed test length effect on bias detection rates is very interesting: When the bias was uniform, the longer test produced a slightly higher bias detection rate of .23, against the rate of the shorter test, of .21. When the bias was nonuniform, the longer test produced a lower detection rate (.37) than the shorter test (.44). The false-positive rates tend to be negatively related to test length. The false-positive rates for the long and short tests were .113 and .157 respectively.

Bias detection rate tends to have a positive relationship with sample size for both uniform and nonuniform bias. For S1, S2 and S3, the bias detection rates were .33, .23 and .19, for uniform bias; and .59, .40 and .22 for nonuniform bias respectively. The false-positive rates displayed a similar trend with .165, .130 and .109 for S1, S2, and S3 respectively.

Considering D1 (items with low discrimination) and D4 (items with high discrimination), both the bias detection rate and the false-positive rate displayed a positive relationship with D1 and D4.

Bias detection and false-positive rates were also estimated for the full chi-square index using the tabled chi-square values at χ = .05 with six degrees of freedom. The bias detection results were worse than those obtained with the P_{ij} values from the no-DIF study. Fewer biased items were detected although the false-positive rates were reduced when the tabled values were used.

Shepard, Camilli, and Williams (1985) and Rudner, Getson, and Knight (1980) both gave the $\chi^2_{i,all}$ the second highest ranking, among the DIF indices in each study, with respect to detecting simulated biased items. Unlike the results in the current study - where the full chi-square performed better with nonuniform bias - Rudner, Getson and Knight (1980) found the result for uniform bias better (a correlation of .86 with simulated bias) than the result for nonuniform bias. As in this study, they found no significant effect of test length on the index's bias detection rate.

The bias detection results obtained for the full chi-square, by Merz and Grossen (1979) were in some ways similar to the results obtained in this study. In both studies the TID performed better than the full chi-square. When the test difficulty increased from .80 to .60 the bias detection rates increased from .33 to .83, when 10% of the items were biased; and from .23 to .83 when 20% of the

items were biased. Similarly, the detection rate increased with difficulty of items in this study.

Seong and Subkoviak (1987) ranked the signed full chi-square better than all the other indices, with respect to bias detection power. Even the IRT-Z index was slightly inferior to the signed chi-square index in their study. This contradicts the results of Shepard, Camilli, and Williams (1985) in which the best performance was recorded for the IRT-Z index.

The Mantel-Haenszel Chi-Square ($\chi^2_{M\text{-}H}$)

Distribution Study

Table 13 gives the means and results from the GLM analysis for the M-H chi-square. Significance was found for sample size for P_{00} and P_{55}. Test length was significant for P_{99}. For P_{90}, S1 and S2 each produced means that were significantly larger than the means for S3. For P_{95}, only S1 produced means that were significantly larger than the means for S3.

Test length tended to have a positive relationship with the means of the P_{99} values: The longer test produced significantly larger means than the shorter test for P_{99}. This test length effect for P_{99} could have been caused by smaller number of observations at each score intervals as there were more score intervals for the longer test.

Table 13

Means by Percentile and Tests of Significance for Different Independent Variables for MANTEL-HAENSZEL CHI-SQUARE

		DISCRIMINATION											
		D1			D2			D3			D4		
Test Length	Sample Size	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}
	S1	1.545	2.315	3.557	1.653	2.497	3.616	1.428	2.181	3.297	1.524	2.386	3.319
TL1	S2	1.647	2.421	3.403	1.380	2.101	3.272	1.519	2.356	3.414	1.734	2.632	3.818
	S3	1.534	2.278	3.296	1.519	2.426	3.786	1.367	2.158	3.179	1.493	2.291	3.502
	S1	1.507	2.250	3.659	1.623	3.550	3.709	1.635	2.431	3.975	1.649	2.453	4.013
TL2	S2	1.464	2.189	4.049	1.577	2.339	3.945	1.505	2.256	4.066	1.489	2.190	3.797
	S3	1.441	2.136	3.518	1.308	1.905	3.370	1.384	2.091	3.450	1.382	2.126	3.664

Significance found for sample size for P_{90} and P_{95}

Significance found for test length for P_{99}

P_{90}
Sample Size Large and Medium > Small

P_{95}
Sample Size Large > Small

P_{99}
Test Length Long > Short

Bias detection power study

The bias detection result is presented in Table 14 while the false-positive results are given in Table 15. The bias detection performance was significantly better on uniform bias than on nonuniform bias. The rates were .96 and .56 for uniform and nonuniform bias respectively. The overall false-positive rate for the M-H chi-square was .095.

When the items were uniformly biased, the detection rates were negatively related to item difficulty. When the items were nonuniformly biased, a much lower detection rate was obtained for items with medium difficulty. This undoubtedly is related to the fact that the intersection of the two ICC's occurred close to the middle of the distribution. The corresponding false-positive rates for low, medium, and high difficulty items were .075, .111, and .094 respectively.

Test length displayed no significant effect on the bias detection rate of the M-H chi-square. However, a very slight increase in detection rate was observed for the longer test. For uniformly biased items, the detection rates were .97 and .96 for the long and short tests respectively. For nonuniformly biased items the detection rates were .57 and .56 respectively for the long and short test. False-positive rate was smaller for the longer test than for the shorter test.

Sample size showed a consistent positive relationship with bias detection rate: For uniform bias, S1, S2 and S3 had rates of 1.00, .99, and .91; while for nonuniform bias the rates were .68,

Table 14

Different Detection Rates (in Proportion) by Levels of Independent Variables for MANTEL-HAENSZEL CHI-SQUARE (at P_{95} Cutoff Value)

		UNIFORM									NONUNIFORM								
		DISCRIMINATION									DISCRIMINATION								
Difficulty Level	Sample Size	D1		D2		D3		D4			D1		D2		D3		D4		
		TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total
Low	S1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.99	.94	.92	.91	.99	.96	.94	.93	.95
	S2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.74	.69	.75	.69	.86	.78	.72	.66	.74
	S3	.99	.90	.96	.94	1.00	1.00	1.00	1.00	.97	.53	.53	.55	.44	.56	.52	.50	.43	.51
	Total	1.00	.97	.99	.98	1.00	1.00	1.00	1.00	.99	.75	.72	.74	.68	.80	.75	.72	.67	.73
Medium	S1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.11	.17	.11	.15	.12	.20	.15	.25	.16
	S2	.99	1.00	.99	1.00	1.00	1.00	1.00	1.00	1.00	.12	.09	.06	.13	.13	.10	.12	.16	.11
	S3	.97	.83	.95	.90	1.00	1.00	.99	.99	.95	.09	.06	.13	.08	.14	.11	.11	.10	.09
	Total	.99	.94	.98	.97	1.00	1.00	1.00	1.00	.98	.11	.11	.10	.12	.13	.14	.13	.17	.12
High	S1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.99	1.00	.99	.98	1.00	.99	.90	.97	.94
	S2	.96	.98	.98	.99	.97	.94	.97	.93	.97	.76	.91	.81	.83	.92	.92	.84	.79	.85
	S3	.78	.83	.81	.77	.80	.73	.73	.77	.80	.61	.68	.85	.59	.62	.72	.60	.67	.67
	Total	.91	.94	.93	.92	.92	.89	.90	.90	.91	.79	.86	.88	.80	.85	.88	.78	.82	.82
	Grand Total	.97	.95	.97	.97	.97	.96	.97	.97	.96	.55	.56	.58	.53	.59	.59	.54	.56	.56

.57 and .42 respectively. Like the detection rate, the false-positive rate also had a positive relationship with sample size.

The bias detection rate of the M-H chi-square, showed no significant effect of item discrimination. The bias detection results when tabled values were used are given in Table 16. These rates are lower by .03 and .12 respectively, for uniform and nonuniform bias, than the rates obtained from the P_{95} cutoff values. The largest difference in bias detection rates for the two bias decision making procedures (i.e. tabled value and P_{95}) was observed in the smallest sample size (S3) used in the study. For uniform and nonuniform biases the differences were .09 and .18 respectively. The smallest detection rate differences for both uniform and nonuniform bias were .00 and .04 respectively. These lowest difference rates were observed for the largest sample size (S1).

Although Hambleton and Rogers (1988) was an empirical data study comparing the M-H chi-square and the ICC area between procedures, some aspects of their findings, with respect to the bias detection power of the M-H chi-square, are worth mentioning here. Among other things, Hambleton and Rogers (1988) observed that of the 16 items in the bias criterion set the M-H chi-square consistently flagged 9 items (detection rate = 56).

Of the seven criterion items not consistently flagged in the two independent comparisons, two items were flagged in one comparison but not the other. These two were classified as false-negative cases. Of the remaining five items, the ICC's of four of them indicated nonuniform bias (i.e. they crossed) to which the M-H

Table 15

False-Positive Rates (in Proportion) by Levels of Independent Variables for MANTEL-HAENSZEL CHI-SQUARE (at P_{95} Cutoff Value)

Difficulty Level	Discrimination	SAMPLE SIZE S1 TL1 First	S1 TL1 Last	S1 TL2 First	S1 TL2 Last	S2 TL1 First	S2 TL1 Last	S2 TL2 First	S2 TL2 Last	S3 TL1 First	S3 TL1 Last	S3 TL2 First	S3 TL2 Last	Total
Low	D1	.077	.078	.230	.087	.074	.063	.175	.078	.062	.059	.137	.028	.096
	D2	.060	.055	.088	.072	.076	.048	.072	.077	.085	.092	.063	.048	.070
	D3	.048	.068	.083	.113	.049	.073	.053	.065	.058	.054	.067	.068	.066
	D4	.066	.074	.097	.080	.075	.070	.045	.048	.071	.076	.055	.067	.069
	Total	.063	.068	.125	.088	.069	.064	.086	.067	..069	.070	.076	.053	.075
Medium	D1	.108	.099	.172	.167	.089	.076	.110	.095	.052	.063	.067	.083	.098
	D2	.085	.093	.120	.145	.076	.084	.107	.138	.101	.096	.070	.085	.100
	D3	.129	.088	.267	.193	.082	.081	.125	.105	.091	.076	.100	.115	.121
	D4	.126	.101	.237	.203	.089	.100	.140	.118	.085	.097	.095	.107	.124
	Total	.112	.095	.198	.177	.084	.085	.121	.114	.082	.083	.083	.098	.111
High	D1	.094	.107	.140	.135	.074	.078	.097	.103	.060	.084	.052	.075	.092
	D2	.063	.111	.112	.115	.076	.062	.123	.122	.095	.108	.085	.075	.095
	D3	.106	.083	.190	.182	.075	.057	.105	.110	.080	.075	.093	.105	.105
	D4	.083	.084	.163	.140	.093	.090	.068	.072	.075	.069	.113	.090	.091
	Total	.087	.096	.151	.143	.080	.072	.098	.100	.078	.084	.086	.086	.096
	Grand Total	.087	.086	.158	.136	.078	.074	.102	.094	.076	.079	.082	.079	.094

Table 16

Different Detection Rates (in Proportion) by Levels of Independent Variables for MANTEL-HAENSZEL CHI-SQUARE (for Tabled Values at Alpha = .05)

		UNIFORM									NONUNIFORM								
		DISCRIMINATION									DISCRIMINATION								
		D1		D2		D3		D4			D1		D2		D3		D4		
Difficulty Level	Sample Size	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total
Low	S1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.96	.87	.90	.86	.98	.92	.90	.85	.91
	S2	1.00	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.54	.64	.59	.51	.76	.60	.50	.44	.64
	S3	.95	.83	.93	.91	1.00	1.00	1.00	1.00	.95	.27	.35	.28	.27	.34	.27	.26	.24	.29
	Total	.98	.94	.98	.97	1.00	1.00	1.00	1.00	.98	.59	.62	.69	.55	.70	.60	.55	.51	.61
Medium	S1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.01	.07	.03	.08	.04	.04	.04	.04	.05
	S2	.99	1.00	.98	1.00	1.00	1.00	1.00	.98	.99	.04	.02	.01	.03	.02	.04	.03	.06	.03
	S3	.92	.85	.90	.85	1.00	.98	.97	.96	.93	.02	.03	.01	.04	.03	.04	.02	.05	.02
	Total	.97	.95	.96	.95	1.00	.99	.99	.98	.97	.02	.04	.02	.04	.03	.04	.03	.05	.03
High	S1	1.00	1.00	1.00	1.00	1.00	.99	.99	1.00	1.00	1.00	.97	.94	.94	.99	.97	.97	.96	.97
	S2	.93	.96	.95	.99	.91	.91	.89	.89	.93	.52	.86	.62	.60	.79	.81	.57	.73	.69
	S3	.52	.72	.65	.63	.56	.48	.51	.54	.58	.29	.44	.27	.40	.45	.49	.35	.53	.40
	Total	.82	.89	.87	.87	.82	.79	.80	.81	.83	.60	.76	.61	.65	.74	.77	.63	74	.69
	Grand Total	.92	.93	.93	.93	.94	.93	.93	.93	.93	.40	.47	.41	.41	.49	.47	.40	.43	.44

chi-square is known to be less sensitive. The ICC's for the remaining item showed that the region on the ability scale which corresponds to the largest differences in probabilities of success, $P(\theta)$, was a region where very few members of one of the groups fell. These results bring out two of the weaknesses of the M-H chi-square index: It is less sensitive to nonuniform bias and in a situation where the ability region with the largest group differences in probabilities of success corresponds to the ability region where very few members of one of the groups fall, the index will behave poorly.

Rogers and Swaminathan (1990) compared the M-H chi-square and the logistic regression DIF procedures. Their findings are very similar to the findings in the current study. They found that for test data generated by the 3-parameter IRT model (which is what was used in the current study), the bias detection rate was .73. This rate, however, was an average taken over different measures of DIF, (.2, .4, .6, .8) an independent variable in their study, which they found to have a positive relationship with DIF detection rate. The current study used only one measure of DIF - an area between value of .6. Thus, if their average detection rate was .73, it is reasonable to assume that for an area between DIF measure of .6, the observed rate would be much higher than .73. This will put the detection rate for uniform bias, with area between value of .6, not very far removed from .96 which was the rate observed in the current study.

The current study also produced M-H chi-square results similar to those obtained by Rogers and Swaminathan (1990) for test length, sample size and item discrimination effects on bias detection rates for both uniform and nonuniform bias.

Sample size effects, similar to the one observed in this study, were also reported by Schulz, Perlman, Rice, and Wright (1989) and Skaggs and Lissitz (1980).

The Mantel-Haenszel Delta (M-H Δ)

Of all the earlier studies, mentioned in Chapter II, that used a Mantel-Haenszel index. Only the study by Skaggs and Lissitz explicitly mention the use of the M-H Δ index in their comparative study.

Distribution study results (M-H Δ)

The results from the GLM analysis are presented in Table 17. These results indicated that the P_{90} of the M-H delta index was significantly affected by sample size, discrimination and sample size by discrimination interaction. The P_{95} and P_{99} values of the index were significantly affected by each of the three independent variables, sample size, test length, and item discrimination, used in the study.

Following the significant sample size by discrimination interaction effect, the simple effect analyses produced the following results: For S1, D4 produced significantly larger means than either D3 or D1. D2 also produced significantly larger means

Table 17

Means by Percentile and Tests of Significance for Different Independent Variables for MANTEL-HAENSZEL DELTA

		DISCRIMINATION											
		D1			D2			D3			D4		
Test Length	Sample Size	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}
	S1	.371	.455	.615	.415	.519	.715	.408	.512	.713	.443	.556	.800
TL1	S2	.553	.686	.953	.606	.763	1.144	.596	.752	1.115	.636	.814	1.198
	S3	.828	1.013	1.459	.859	1.092	1.716	.896	1.117	1.656	.963	1.195	1.817
	S1	.377	.470	.593	.421	.529	.678	.385	.496	.634	.430	.552	.710
TL2	S2	.570	.701	.874	.570	.735	.987	.595	.780	1.007	.691	.879	1.137
	S3	.829	1.042	1.371	.908	1.184	1.741	.882	1.123	1.527	.978	1.250	1.717

Significance found for sample size, discrimination and sample size discrimination intercaction for P_{90}
Significance found for sample size, test length and discrimination for P_{95} and P_{99}

P_{90} Simple Effects
Large Sample Size D4 > D3, D1 and D2 > D1
Medium Sample Size D4> D3, D2, D1
Small Sample Size D4 > D3, D2, D1 and D3, D2 > D1
For D1 D2 D3, D4 Small Sample Size > Medium > Large

P_{95}
Sample Size Small > Medium > Large
Test Length Short > Long
Discrimination D4 > D3, D2 > D1

P_{99}
Sample Size Small > Medium > Large
Test Length Long > Short
Discrimination D4 > D3, D1 and D2, D3 > D1

than D1. For S2, D4 produced significantly larger means than either D3, D2 or D1. For S3, D4 produced significantly larger means that either D3, D2 or D1. Also D3 and D2 each produced significantly larger means than D1. Considering the two main discrimination levels, D1 (low discrimination) and D4 (high discrimination), it is clear that at each sample size level D4 consistently produced significantly larger means than D3 and D1.

At each of the four discrimination levels, D1, D2, D3 and D4, sample size was negatively related to the P_{90} value of the M-H delta.

For P_{95}, sample size was observed to be negatively related to the size of the mean of the M-H delta. Test length was also negatively related to the size of the M-H delta mean for P_{95}. In terms of D1 and D4, discrimination indicated a positive relationship with the means of the index for P_{95}.

For P_{99}, sample size again displayed a negative relationship with the index. On the other hand test length was found to be positively related to the index. Discrimination is again observed to have a positive relationship with the index when only D1 and D4 are considered.

The main implication for these results is that whenever the M-H delta index is used, the user must remember that the bias decision method must take into account the sample size, test length and item discrimination in the particular situation being studied for bias. For example, if a baseline statistic is used to make the

bias decision, the sample size used in the baseline study must be the same as that used in the bias study.

Power study results (M-H Δ)

Bias detection results and false-positive rates, for the M-H delta, are presented in Table 18 and Table 19 respectively. The results show that the M-H delta index performed significantly better on items with uniform bias than on items with nonuniform bias. For uniform bias the detection rate was .92 while the rate for nonuniform bias was .44. The overall false-positive rate was .081.

The bias detection rate, for uniformly biased items, displayed a negative relationship with difficulty level. There was almost a perfect detection rate (.99) for low difficulty items with uniform bias. The rate dropped for items with medium and high difficulties. When the items were nonuniformly biased, the detection rates were similar for biased items with low and high difficulties respectively. Medium difficulty items with uniform bias were very poorly detected at a rate of .05. The false-positive rate tends to have a relationship with item difficulty that is similar to that observed between item difficulty and bias detection rate.

Bias detection rates, for both uniform and nonuniform bias, displayed no meaningful differences between TL1 and TL2.

For both uniform and nonuniform bias, bias detection rates and false positive rates were higher for larger samples than for smaller samples.

Table 18

Different Detection Rates (in Proportion) by Levels of Independent Variables for MANTEL-HAENSZEL DELTA (at P_{95} Cutoff Value)

		UNIFORM									NONUNIFORM								
		DISCRIMINATION									DISCRIMINATION								
		D1		D2		D3		D4			D1		D2		D3		D4		
Difficulty Level	Sample Size	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total
Low	S1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.98	.89	.87	.86	.94	.84	.87	.84	.89
	S2	1.00	1.00	1.00	.99	1.00	1.00	1.00	1.00	1.00	.63	.59	.70	.63	.76	.67	.56	.42	.62
	S3	.96	.88	.91	.88	1.00	1.00	1.00	1.00	.96	.44	.42	.45	.32	.39	.33	.37	.28	.37
	Total	.99	.96	.97	.96	1.00	1.00	1.00	1.00	.99	.68	.63	.67	.60	.70	.61	.60	.51	.63
Medium	S1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.06	.09	.06	.07	.05	.05	.05	.07	.06
	S2	.99	.97	.98	.99	1.00	.99	1.00	.99	.99	.07	.07	.02	.04	.03	.05	.03	.07	.05
	S3	.94	.83	.86	.81	1.00	.95	.96	.92	.91	.03	.04	.04	.01	.02	.04	.02	.03	.03
	Total	.98	.93	.95	.93	1.00	.98	.99	.97	.93	.04	.07	.04	.04	.03	.05	.03	.06	.05
High	S1	.99	1.00	1.00	1.00	.99	.99	.99	.97	.99	.97	.97	.92	.93	.98	.94	.91	.88	.94
	S2	.95	.96	.90	.96	.91	.82	.86	.76	.89	.56	.85	.62	.59	.74	.72	.46	.54	.64
	S3	.61	.74	.69	.64	.60	.66	.53	.63	.64	.31	.48	.34	.23	.39	.31	.27	.38	.34
	Total	.85	.90	.86	.87	.83	.82	.79	.79	.84	.61	.77	.63	.58	.70	.66	.55	.60	.64
	Grand Total	.94	.93	.93	.92	.94	.93	.93	.92	.92	.44	.49	.45	.41	.48	.44	.39	.39	.44

Table 19

False-Positive Rates (in Proportion) by Levels of Independent Variables for M-H DELTA (at P_{95} Cutoff Value)

Difficulty Level	Discrimination	SAMPLE SIZE S1 TL1 First	S1 TL1 Last	S1 TL2 First	S1 TL2 Last	S2 TL1 First	S2 TL1 Last	S2 TL2 First	S2 TL2 Last	S3 TL1 First	S3 TL1 Last	S3 TL2 First	S3 TL2 Last	Total
Low	D1	.116	.120	.260	.148	.116	.115	.213	.130	.121	.109	.177	.083	.142
	D2	.086	.138	.100	.158	.103	.146	.077	.070	.110	.176	.088	.128	.115
	D3	.131	.098	.173	.143	.126	.106	.135	.090	.154	.085	.167	.097	.126
	D4	.124	.157	.157	.175	.144	.131	.115	.118	.139	.161	.135	.130	.141
	Total	.107	.128	.173	.159	.122	.125	.135	.102	.131	.133	.117	.110	.131
Medium	D1	.064	.050	.123	.105	.051	.034	.055	.055	.029	.031	.030	.040	.056
	D2	.039	.053	.058	.087	.031	.048	.087	.122	.036	.052	.022	.042	.057
	D3	.077	.080	.198	.098	.049	.036	.065	.048	.038	.020	.055	.042	.067
	D4	.070	.040	.128	.113	.036	.046	.058	.060	.036	.040	.047	.043	.060
	Total	.063	.056	.127	.101	.042	.041	.066	.071	.035	.036	.039	.042	.060
High	D1	.060	.065	.093	.080	.041	.039	.051	.057	.032	.044	.028	.033	.052
	D2	.028	.073	.050	.083	.033	.033	.111	.098	.031	.049	.020	.037	.063
	D3	.052	.036	.115	.088	.042	.025	.062	.043	.045	.029	.053	.047	.053
	D4	.042	.038	.070	.068	.043	.040	.033	.030	.021	.026	.040	.038	.040
	Total	.045	.053	.082	.080	.040	.034	.064	.057	.032	.037	.035	.039	.052
	Grand Total	.072	.079	.127	.113	.068	.067	.088	.077	.056	.069	.064	.064	.081

Item discrimination does not seem to affect bias detection rate or false-positive rate of the M-H delta in any way.

The M-H chi-square and the M-H delta bias detection results displayed identical patterns in their relationships with the different independent variables; although the rates obtained for the M-H delta were lower on every level of the independent variables. The pattern of the relationships between the independent variables, test length, sample size, and discrimination, and the false-positive rate were identical for both the M-H chi-square and M-H delta. However, the false-positive rates at corresponding levels of the three independent variables (test length, samples size, and discrimination) were lower for the M-H delta at every level. The false-positive rates at the different levels of item difficulty displayed different patterns for the M-H chi-square, the false-positive rates for low, medium and high difficulties were .075, .111, and .094 respectively. For the M-H delta, on the other hand, a negative relationship was observed.

The study by Skaggs and Lissitz (1988) reported a slight superiority, in terms of bias detection consistency, of the M-H chi-square over the M-H delta in the three sample size conditions examined in their study. This is in line with the findings in this study where the M-H chi-square returned a better detection rate in all the experimental conditions.

The Logistic Regression (LR) Index

As was mentioned in the methodology section, two separate procedures (one for uniform DIF and one for nonuniform DIF) were used for the LR procedure. The means and GLM analysis results for the uniform DIF procedure (LR1) are presented in Table 20 while those for the nonuniform DIF procedure (LR2) are presented in Table 21. The bias detection results of the two LR procedures are presented together in Table 22. The false-positive analysis results are presented separately in Table 23, for LR1; and in Table 24, for LR2. The bias detection results when tabled values (at alpha = .05) are presented in Table 25.

Distribution study results (LR)

From Table 20, the only significant result observed for LR1 is for test length for P_{99}. Longer tests produced significantly larger means for P_{99} than shorter tests.

The results reported in Table 21 for LR2, indicate significance found for sample size and the sample size by test length interaction for P_{90}; plus a significant test length effect for P_{95} and P_{99}. Following the significant sample size by test length interaction effect, simple effects analyses were undertaken. For the long test, no significant differences were observed for P_{90}, between the sample size levels. However, for the short test, S2 produced means that were significantly larger than those for S1 for P_{90}. No significant differences were observed between the long and short test means at each sample size level.

Table 20

Means by Percentile and Tests of Significance for Different Independent Variables for LOGISTIC REGRESSION UNIFORM DIF

		DISCRIMINATION											
		D 1			D 2			D 3			D 4		
Test Length	Sample Size	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}
	S1	1.675	2.454	3.968	1.774	2.553	3.961	1.824	2.645	4.224	1.811	2.589	4.407
TL1	S2	1.719	2.448	4.489	1.861	2.715	4.371	1.736	2.602	4.501	1.747	2.541	4.308
	S3	1.819	2.621	4.268	1.634	2.392	4.039	1.760	2.541	4.186	1.798	2.596	4.304
	S1	1.702	2.499	3.834	1.848	2.716	3.853	1.588	2.392	3.513	1.713	2.617	3.574
TL2	S2	1.914	2.703	3.626	1.654	2.476	3.559	1.757	2.683	3.776	2.034	2.992	4.306
	S3	1.853	2.730	3.980	1.898	2.906	4.418	1.748	2.580	3.771	1.886	2.739	4.121

Significance found for test length at P_{99}

P_{99} Test Length Long > Short

Table 21

Means by Percentile and Tests of Significance for Different Independent Variables for LOGISTIC REGRESSION NONUNIFORM DIF

		DISCRIMINATION											
		D1			D2			D3			D4		
Test Length	Sample Size	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}	P_{90}	P_{95}	P_{99}
	S1	1.699	2.452	4.086	1.914	2.781	4.730	1.851	2.816	4.804	1.885	2.794	4.931
TL1	S2	1.737	2.511	4.606	1.782	2.580	4.740	1.660	2.536	4.367	1.767	2.682	4.571
	S3	1.757	2.633	4.454	1.896	2.751	4.807	1.882	2.767	4.772	1.906	2.851	5.129
	S1	1.809	2.850	3.963	1.804	2.613	3.699	1.788	2.646	3.800	1.878	2.789	3.818
TL2	S2	2.014	2.932	4.073	1.800	2.652	3.632	1.922	2.749	4.273	2.237	3.245	4.708
	S3	1.988	3.007	4.555	1.960	2.819	3.962	1.959	2.840	4.050	1.965	2.868	4.271

Significance found for sample size and sample size-test length interaction for P_{90}

Significance found for test length for P_{95} and P_{99}

P_{90} Simple Effects
Test Length - Large Sample Size No Difference
Short Sample Size Medium < Large
Sample Size - Large No Difference in Test Length
Medium No Difference in Test Length
Short No Difference in Test Length

P_{95} Test Length Short > Long

P_{99} Test Length Long > Short

Table 22

Different Detection Rates (in Proportion) by Levels of Independent Variables for LOGISTIC REGRESSION (UNIFORM AND NONUNIFORM) (at P_{95} Cutoff Value)

		UNIFORM									NONUNIFORM								
		DISCRIMINATION									DISCRIMINATION								
		D1		D2		D3		D4			D1		D2		D3		D4		
Difficulty Level	Sample Size	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total
Low	S1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	S2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.92	.95	1.00	.97	.98	.95	.98	.99	.97
	S3	.99	.91	.95	.95	1.00	1.00	1.00	1.00	.99	.64	.79	.81	.70	.77	.75	.81	.74	.76
	Total	1.00	.97	.98	.98	1.00	1.00	1.00	1.00	.99	.85	.91	.94	.91	.92	.90	.93	.91	.91
Medium	S1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.99	1.00
	S2	.99	1.00	.99	1.00	1.00	1.00	1.00	1.00	1.00	.91	.91	.92	.89	.98	.95	.98	.88	.95
	S3	.97	.95	.87	.94	1.00	.99	.98	.99	.96	.72	.65	.68	.64	.76	.69	.74	.68	.70
	Total	.99	.96	.95	.98	1.00	1.00	.99	1.00	.98	.88	.85	.87	.84	.93	.88	.91	.85	.88
											Logistic Regression #1 (uniform index)								
High	S1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.99	.99	.99	.98	1.00	.99	1.00	.96	.92
	S2	.98	.98	.96	.99	.97	.96	.97	.92	.97	.75	.91	.75	.82	.90	.92	.79	.79	.74
	S3	.81	.84	.83	.81	.84	.73	.73	.79	.80	.51	.64	.47	.55	.62	.71	.56	.66	.48
	Total	.93	.94	.93	.93	.94	.90	.90	.90	.92	.75	.85	.74	.79	.84	.87	.82	.80	.81
	Grand Total	.97	.96	.96	.96	.98	.97	.96	.97	.96	.84	.87	.85	.85	.90	.88	.89	.85	.87

Table 23

False-Positive Rates (in Proportion) by Levels of Independent Variables for LR1 (at P_{95} Cutoff Value)

Difficulty Level	Discrimination	SAMPLE SIZE S1 TL1 First	S1 TL1 Last	S1 TL2 First	S1 TL2 Last	S2 TL1 First	S2 TL1 Last	S2 TL2 First	S2 TL2 Last	S3 TL1 First	S3 TL1 Last	S3 TL2 First	S3 TL2 Last	Total
Low	D1	.084	.074	.235	.088	.069	.074	.170	.077	.061	.058	.135	.043	.097
	D2	.063	.064	.080	.075	.077	.056	.077	.070	.086	.090	.070	.052	.072
	D3	.044	.062	.085	.118	.061	.073	.057	.080	.067	.060	.082	.078	.072
	D4	.071	.081	.093	.088	.074	.072	.057	.058	.059	.073	.070	.067	.072
	Total	.068	.070	.123	.092	.070	.069	.090	.071	.069	.070	.089	.060	.078
Medium	D1	.085	.087	.165	.157	.079	.068	.100	.078	.052	.047	.070	.077	.090
	D2	.082	.083	.115	.125	.073	.072	.087	.122	.092	.087	.062	.078	.080
	D3	.113	.077	.233	.158	.073	.074	.122	.095	.082	.062	.097	.095	.106
	D4	.108	.092	.207	.182	.089	.084	.127	.110	.073	.082	.082	.112	.112
	Total	.097	.083	.180	.156	.079	.075	.109	.101	.074	.070	.078	.091	.100
High	D1	.079	.091	.115	.098	.067	.077	.088	.093	.055	.056	.055	.070	.079
	D2	.056	.086	.098	.103	.071	.053	.112	.098	.078	.092	.072	.060	.082
	D3	.078	.075	.157	.148	.061	.060	.080	.102	.080	.073	.090	.102	.092
	D4	.084	.084	.133	.117	.076	.075	.048	.058	.063	.066	.087	.080	.081
	Total	.074	.084	.126	.117	.069	.066	.082	.088	.069	.072	.076	.078	.083
	Grand Total	.080	.079	.143	.122	.073	.070	.094	.087	.071	.071	.081	.076	.087

Table 24

False-Positive Rates (in Proportion) by Levels of Independent Variables for LR2 (at P_{95} Cutoff Value)

Difficulty Level	Discrimination	SAMPLE SIZE S1 TL1 First	S1 TL1 Last	S1 TL2 First	S1 TL2 Last	S2 TL1 First	S2 TL1 Last	S2 TL2 First	S2 TL2 Last	S3 TL1 First	S3 TL1 Last	S3 TL2 First	S3 TL2 Last	Total
Low	D1	.065	.077	.215	.093	.071	.069	.153	.072	.049	.046	.103	.053	.090
	D2	.064	.070	.082	.065	.084	.070	.098	.083	.073	.075	.072	.063	.075
	D3	.048	.051	.098	.118	.071	.064	.070	.080	.073	.071	.105	.067	.076
	D4	.067	.076	.077	.083	.068	.082	.067	.072	.072	.072	.082	.078	.075
	Total	.061	.069	.118	.090	.074	.071	.097	.077	.067	.066	.091	.065	.079
Medium	D1	.078	.066	.088	.085	.065	.063	.048	.067	.062	.063	.047	.040	.064
	D2	.061	.054	.092	.072	.066	.078	.087	.065	.074	.068	.063	.065	.070
	D3	.050	.040	.112	.103	.074	.049	.083	.087	.080	.084	.083	.082	.077
	D4	.062	.069	.087	.075	.074	.063	.063	.062	.075	.067	.068	.070	.070
	Total	.063	.057	.095	.084	.070	.063	.070	.070	.073	.071	.065	.064	.071
High	D1	.099	.086	.093	.095	.076	.081	.080	.062	.059	.055	.065	.062	.076
	D2	.070	.066	.047	.108	.066	.079	.105	.090	.067	.098	.057	.093	.079
	D3	.068	.056	.123	.112	.072	.067	.085	.080	.083	.074	.105	.097	.085
	D4	.066	.082	.098	.098	.089	.066	.050	.063	.059	.069	.087	.095	.077
	Total	.076	.073	.090	.103	.076	.073	.080	.074	.067	.074	.179	.087	.079
	Grand Total	.067	.066	.101	.092	.073	.069	.082	.074	.069	.070	.112	.072	.076

Table 25

Different Detection Rates (in Proportion) by Levels of Independent Variables for LOGISTIC REGRESSION (for Tabled Values at Alpha = .05)

		UNIFORM									NONUNIFORM								
		DISCRIMINATION									DISCRIMINATION								
Difficulty Level	Sample Size	D1		D2		D3		D4			D1		D2		D3		D4		
		TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total	TL1	TL2	TL1	TL2	TL1	TL2	TL1	TL2	Total
	S1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Low	S2	1.00	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.84	.95	.96	.94	.98	.91	.98	.97	.95
	S3	.97	.89	.96	.96	1.00	1.00	1.00	1.00	.97	.50	.66	.67	.59	.65	.64	.73	.65	.64
	Total	.99	.96	.99	.99	1.00	1.00	1.00	1.00	.99	.78	.87	.88	.84	.88	.85	.90	.87	86
	S1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.99	1.00	1.00
Low	S2	1.00	1.00	.99	1.00	1.00	1.00	1.00	1.00	1.00	.84	.89	.87	.81	.92	.87	.97	.86	.88
	S3	.96	.91	.95	.97	1.00	.99	.98	.99	.97	.48	.61	.58	.48	.67	.59	.64	.58	.58
	Total	.99	.99	.98	.99	1.00	1.00	.98	1.00	.99	.77	.83	.82	.76	.86	.82	.87	.81	.83
											Logistic Regression (Uniform index)								
	S1	1.00	1.00	1.00	1.00	1.00	.99	.99	1.00	1.00	1.00	.99	.87	.93	.99	.98	.97	.96	.96
High	S2	.95	.98	.96	.98	.94	.94	.99	.92	.95	.57	.89	.65	.62	.86	.84	.62	.75	.74
	S3	.63	.77	.71	.84	.66	.56	.61	.68	.68	.34	.52	.30	.42	.47	.55	.37	.53	.44
	Total	.86	.92	.89	.91	.87	.83	.86	.87	.88	.67	.80	.61	.66	.77	.82	.65	.75	.71
	Grand Total	.95	.96	.95	.96	.96	.94	.95	.96	.95	.74	.83	.77	.75	.84	.83	.81	.81	.80

For P_{95} and P_{99}, the short test produced significantly larger means than the long test. For P_{99}, however, the long test produced significantly larger means than the shorter test.

Power study results (LR)

First of all, it must be noted that for each of the six biased items results were reported in Table 22. It should be noted that results for LR1 were used for all three uniformly biased items and for nonuniformly biased items with high difficulty. With LR2, results were reported for nonuniformly biased items at the low and medium difficulty levels. This was done because both indices are available when bias decisions are made. The results for false-positive rates are presented in Table 23 and Table 24 for LR1 and LR2 respectively.

In general the LR procedure performed significantly better in detecting uniform bias than in detecting nonuniform bias items. The bias detection rates were .96 and .87 for uniform and nonuniform bias, respectively, with the P_{95} cutoff value. When tabled values (at alpha = .05) were used the detection rates for uniform and nonuniform bias were .95 and .80 respectively. The false-positive rates (with P_{95}) were .087 and .076 for uniform and nonuniform bias respectively.

For both uniform and nonuniform biases the detection rate of the LR procedure increased as item difficulty decreased. The highest false-positive rate was observed for items with medium

difficulty while the lowest false-positive rate was observed for items with low difficulty.

Test length showed no effects on the bias detection rates of both uniform and nonuniform biases when the P_{95} cutoff was used. The tabled values (at α = .05) showed a very slight nonsignificant negative trend, for test length on bias detection. A negative trend was observed for the test length effect on false-positive rate.

Both the P_{95} cutoff and the tabled value (at alpha = .05) results produced positive relationships between sample size and bias detection rate for both uniform and nonuniform bias. The false-positive rates for S3 and S2 were equal at .078. The false positive rate was higher for S1, than those observed for S3 and S2, by about .02. No discrimination effects were observed, for uniform or nonuniform bias or false-positive rates, in the results from either the P_{95} cutoff or the tabled values (at alpha = .05).

The LR index was proposed by Rogers and Swaminathan (1989) out of their concern for one major weakness in the Mantel-Haenszel DIF procedure. The weakness was the fact that the M-H DIF procedures seemed to be less sensitive to nonuniform DIF - in particular, nonuniform DIF with no difference in the item's b-value for the studied groups. In Rogers and Swaminathan (1989, 1990) the logistic regression produced results that indicated this particular weakness of the M-H procedure, is appreciably handled by the logistic regression procedure. In the current study, much as both the M-H chi-square and the logistic regression procedures performed equally well on items with uniform bias, when the bias was nonuniform, with

no b-value difference between the studied groups, the M-H chi-square and the LR procedures produced detection rates of .56 and .87 respectively.

The summary of findings, limitation of this study, and suggestions for further studies will be presented in the next chapter.

CHAPTER VI

Conclusion

This chapter will contain the following sections: Summary of findings, limitations of study, and suggestions for further research.

Summary of Findings

The method of presentation of the summary will be the same as that used in the result presentation section: The summary will be presented first for the distribution study followed by the presentation for the power study.

Distribution study

Three independent variables were used in this study. These were sample size, test length and item discrimination. However, for indices based on the IRT model, the sample size independent variable was omitted. Thus, only the non-IRT indices were examined for sample size effect. Of these, four were observed to have sample size effects. These were the transformed item difficulty (TID), the Mantel-Haenszel chi-square (M-H chi-square), the Mantel-Haenszel delta (M-H delta), and the Logistic regression (LR) for nonuniform DIF. Test length was found to affect all but one of the eight indices. The only index not affected by test length was Lord's chi-

square. Discrimination effects were observed for the SOS2, M-H delta, and the full chi-square indices.

Power study

The summary is presented in terms of the performance of the indices with respect to their detection rates for uniform and nonuniform bias and their false-positive rates. This being the case, it is not possible to rank order the indices with respect to their performance in the power study. The results clearly show that there is no index, in the study that, that would occupy the same hierarchical position in each of the three areas of performance - uniform and nonuniform bias detection rates and false-positive rates.

Uniform bias detection

The results show that the logistic regression, the M-H chi-square, the M-H delta and Lord's chi-square performed very well on uniformly biased items. Each of these had a detection rate greater than .90; with Lord's chi-square returning a perfect 1.00, detection rate. The TID performed averagely; with a detection rate of .62. The SOS2 and the full chi-square did not perform very well in this study. The detection rates for uniformly biased items were .32 and .25 for the SOS2 and the full chi-square respectively.

Nonuniform bias detection

For the detection of nonuniform bias, Lord's chi-square stands uniquely alone in the greater then .90 rate class. Again, Lord's chi-square gave a perfect, 1.00, detection rate. The next best performance on nonuniform bias detection was observed for the logistic regression index which produced a detection rate of .87 in this class. Three indices, the full chi-square, the M-H chi-square, and the M-H delta produced detection rates between .40 and .60; with the M-H chi-square leading with a rate of .56. The TID and the SOS2 performed very poorly with rates of .05 and .01 respectively.

False-positive rates

The false-positive identification results show the SOS2 with the best performance in this class. It produced the lowest rate of .053. The M-H delta and Lord's chi-square had the next best performance - each with a false-positive rate of .081 - to the SOS2. Not far behind the M-H delta and Lord's chi-square was the logistic regression index with a false-positive rate of .082. The M-H chi-square and the full chi-square had false-positive rates of .094 and .135 respectively.

From the results summary above, Lord's chi-square certainly appears to perform very well. Its detection rates were the best and its percentiles were not affected by either test length or item discrimination. When tabled values were used, Lord's chi-square produced bias detection results similar to the one for P_{95} reported in Table 5. The tabled value results also produced an improvement

in terms of reduced false-positive rates. However, the false-positive rates (Table 6) were still higher than those produced by the SOS2. In retrospect, one would like to know how well it might do with smaller sample sizes.

Of the other indices, the Mantel-Haenszel and the logistic regression indices seemed the best. The full chi-square had a number of problems. Its tabled values were not at all useful for detection of bias. The TID was somewhat better but does not have a significant test associated with it. One would need to rely on baseline studies if one were to use it. Even so, since sample size and test length had effects on the TID percentiles; care must be taken to make sure that the baseline study sample size and test length are the same as those used in the bias study.

For uniform bias, either Mantel-Haenszel chi-square or logistic regression would be recommended while for nonuniform bias logistic regression would be appropriate. It is interesting to note that Lord's chi-square was effective for detecting either kind of bias.

It is a well established fact that sample size is related positively to chi-square values. In the study, however, even though samples as large as 1,600 subjects and as small as 150 subjects were used, for each of the chi-square indices, the observed values were considerably lower than tabled values. Of course, these were conditions where no bias was present except those that might have been randomly induced in data generation. The observed and tabled value relationship, mentioned above, could be explained either in

terms of a high statistical power (i.e. a high probability of rejecting the "no bias" null hypothesis, when the alternative - or "bias present" hypothesis is true) or perhaps it is those instances where bias is truly present that larger sample sizes allow a much easier identification of biased items. Certainly, the proportion of biased items detected was greater for large sample sizes for the full chi-square, the M-H chi-square and the LR chi-square indices. Because of the overall high detection rates this phenomenon could not be observed with Lord's chi-square.

It should be noted that in this study, all the obtained values for the chi-square indices for P_{95} were smaller than the corresponding tabled values (at alpha = .05). To this end, it would appear safer to use the tabled values whenever bias studies, involving chi-square indices, are done. This will avoid the apparent conservative nature of the P_{95} baseline value.

Limitations of Study

The manipulation of independent variables, made possible by the use of a simulated study, may be considered the greatest limitation in this study. However carefully a simulation study is planned, it will always lack the human aspects in real life data studies. The complex nature of the human mind, which is normally called into action in a problem solving situation cannot be modeled in simulation studies of any form. In simulation studies, only a few of the factors (some of those we know of) are built into the

simulation model. Some of the factors we know of, but believe their influence will be minimal in the study; and those other factors we do not know about, are all not included in our study. Thus, there is much more involved in real life data studies than we look at in simulated data studies.

It is also possible that the results in a simulated data study are purely a reflection of the restriction applied to the data. There is a possibility that if this study is repeated with different levels of the independent variables the results may be altered considerably. All the restrictions imposed on the different parameters used in this study may have affected the results obtained to some extent. However, if for a series of similar studies, with different restrictions placed on the same parameters, results that are similar to those in this study are found, a lot more confidence could then be placed on the obtained results of this study.

Suggestions for Further Studies

From the end part of the last section, it follows, that two forms of this study should be recommended for further research. In the first place, although the result of this study could be generalized to similar data sets, I would like to see this study replicated with all the parameters identical, in forms and measures, to this study. Secondly, I would also like to see the

study replicated on data sets obtained from the use of different values than the parameters used in this study.

In the distribution study, the independent variables were observed to affect the three percentile values of the indices in varying combinations. Such results are indications of problems with the index concerned. If, for example, it can be established beyond all doubts that sample size affects the percentile values of our index, then using any of the percentiles in bias decision making must take into account the particular sample size used in the bias study. If, on the other hand, the index claims to have a known statistical test of significance, for example, the M-H chi-square, which does not take sample size into account, the sample size effect may be distorting the obtained bias result in a very serious way. Since longer test are known to produce better estimates of a person's true score on the ability measured by the test, it stands to reason that a bias detection index that displays test length effect may be confounding bias with group ability. A similar argument could be presented for confounding discrimination with bias.

From the results obtained from the distribution studies four, seven and three of the indices were affected by sample size, test length and item discrimination respectively. Other research results are needed, for the effects of these independent variable, to verify the findings in this study. For if these results are true, beyond all reasonable doubt, this puts a big question mark on the validity and reliability of such indices.

In this study, the effect of sample size on the IRT DIF procedures was inconclusive because of the convergence problem with LOGIST. If computer programs, that avoid this limitation of LOGIST. for example, BILOG, could be accessed, it is worth examining the effect of sample size on the SOS indices and Lord's chi-square.

REFERENCES

Ackerman, T.A. (1991). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Il.

Anastasi, A. (1976). Psychological testing (4th ed.) Toronto: Macmillan.

Angoff, W.H. (1972). A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu, Hi.

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.) Handbook of methods for detecting test bias (pp. 96-116). Baltimore: Johns Hopkins University Press.

Angoff, W.H., & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Baker, F.B. (1981). A criticism of Scheuneman's item bias technique. Journal of Educational Measurement, 18, 59-62.

Burgon, M.S., & Lissak, R.I. (1981). Detecting item bias using the ICC3 and the B and C statistics. Paper presented at the annual meeting of the American Psychological Association. Los Angeles, Ca.

Burrill, L.E. (1982). Comparative studies of item bias methods. In R.A. Berk (Ed.), Handbook of methods for detecting test bias. (pp. 161-179). Baltimore: Johns Hopkins University Press.

Camilli, G. (1979). A critique of the chi-square for assessing item bias. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder.

Dorans, N.J. & Kulick, E. (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.

Draba, R.E. (1977). The identification and interpretation of item bias. (Memorandum No. 26). Chicago: Chicago University Department of Education, Statistical Laboratory.

Durovic, J.J. (1975). Test bias: An objective definition for test items. Paper presented at the annual meeting of the Northeastern Educational Research Association. Ellenville, N.Y.

Echternacht, G. (1974). A quick method for detecting test bias. Educational and Psychological Measurement, 34, 271-280.

Green, D.R., & Draper, J.F. (1972). Exploratory studies of bias in Achievement Tests. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, Ca.

Hambleton, R.K., & Rogers, H.J. (1988). Detecting biased test items: Comparison of the IRT Area and Mantel-Haenszel methods. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, La.

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

Hills, J.R. (1989). Screening for potentially biased items in testing programs. Educational Measurement Issues and Practice, 8(4), 5-11.

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

Hulin, C.L., Drasgow, F., & Komocar, J. (1982). Application of item response theory to analysis of Attitude Scale Translation. Journal of Applied Psychology, 67, 818-825.

Ironson, G.H. (1982). Use of chi-square methods and latent trait approaches for determining item bias. In R.A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 117-160). Baltimore: Johns Hopkins University Press.

Ironson, G.H., Homan, S., Willis, R., & Signer, B. (1984). The validity of item bias techniques with math word problems. Applied Psychological Measurement, 8, 391-396

Ironson, G.H., & Subkoviak, M.J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 16, 209-225.

Lei, H., & Skinner, H.A. (1982). What difference does language make? Structural analysis of the personality research form. Multivariate Behavioral Research, 17, 33-46.

Linn, R.L., Levine, M.V., Hastings, C.N., & Wardrop, J.L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Linn, R.L., & Harnisch, D.L. (1981). Interaction between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

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

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

Marascuilo, L.A., & Slaughter, R.E. (1981). Statistical procedures for identifying possible sources of item bias based on chi-square statistics. Journal of Educational Measurement, 18, 229-248.

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

Merz, W.R. (1973). Factor analysis as a technique in analyzing test bias. Paper presented at the annual meeting of the California Educational Research Association, Los Angeles, Ca.

Merz, W.R. (1976). Estimating bias in test items utilizing principle component analysis and the general linear solution. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Ca.

Merz, W.R., & Grossen, N.E. (1979). An empirical investigation of six methods for examining test item bias. (Tech. Rep. No. Grant NIE-6-78-0067). Sacramento: California State University, National Institute of Education.

Nunnally, J.C. (1967). Psychometric theory. New York: Mcgraw-Hill.

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

Rock, D.A., & Werts, C.E. (1979). Construct validity of the SAT across population: An empirical confirmatory study. (Tech. Rep. No. PR-79-2). New Jersey: Educational Testing Services.

Rudner, L.M. (1977). An approach to biased item identification using latent trait measurement theory. Paper presented at the annual meeting of the American Educational Research Association. New York, NY.

Rudner, L.M., & Convey, J.J. (1978). An evaluation of select approaches for biased item identification. paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada.

Rudner, L.M., Getson, P.R., & Knight, D.L. (1980). Biased item detection techniques. Journal of Education Statistics, 5, 213-233.

Rudner, L.M., Getson, P.R., & Knight, D.L. (1980). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17, 1-10.

Schulz, E.M., Perlman, C., Rice, W.K. & Wright, B.D. (1989). An empirical comparison of Rasch and Mantel-Haenszel procedures for assessing item bias. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, Ca.

Seong, T.J., & Subkoviak, M.J. (1987). A comparative study of recently proposed item bias detection methods. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

Shepard, L., Camilli, G., & Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

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

Scheuneman, J. (1982). A posteriori analyses of biased items. In R.A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 180-198). Baltimore: Johns Hopkins University Press.

Scheuneman, J. (1979). A new method for assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Shaw, D.G., Huffman, M.D., & Haviland, M.G. (1987). Grouping continuous data in discrete interval: information loss and recovery. Journal of Educational Measurement, 24, 167-173.

Skaggs, G., & Lissitz, R.W. (1988). Consistency of selected item bias indices: Implications of another failure. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, La.

Spray, J.A. (1989). Performance of three conditional DIF statistics in detecting differential item functioning on simulated test. (ACT Report No. 89-7). Iowa: American College Testing Program.

Spray, J.A., & Carlson, J. (1986). Comparison of loglinear and logistic regression models for detecting changes in proportions. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Ca.

Stricker, L.J. (1982). Identifying test items that perform differentially in population subgroups: A partial correlation index. Applied Psychological Measurement, 6, 216-273.

Subkoviak, M.J., Mack, J.S., Ironson, G.H., & Craig, R.D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. Journal of Educational Measurement, 21, 49-58.

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

Swaminathan, H., & Rogers, H.J. (1989). Detecting item bias using logistic regression. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Ca.

Tittle, C.K. (1982). Use of judgemental methods in item bias studies. In R.A. berk (Ed.), Handbook of methods for detecting test bias (pp. 31-63). Baltimore: Johns Hopkins University Press.

Traub, R.E. (1983). A priori considerations in choosing item response model. In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver: Educational Research Institute of British Columbia.

Vaillancourt, R. (1984). IRT bias detection techniques compared with classical item analysis as applied to Anglophones and Francophones. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, La.

Vander Flier, H., Mellenbergh, G.J., Ader, H.J., & Wijn, M. (1984). An iterative item bias detection method. Journal of Educational Measurement, 21, 131-145.

Veale, J.R., & Foreman, D.I. (1983). Assessing cultural bias using foil response data: Cultural variation. Journal of Education Measurement, 20, 249-258.

Welch, C.J., Ackerman, T.A., Doolittle, A.E., & Hurley, J. (1987). An examination of statistical procedures for detecting cross-cultural differential item performance. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

Wright, B.D., Mead, R.J., & Draba, R. (1976). Detecting and correcting test item bias with a logistic response model. (Tech. Rep. No. 22). Chicago: Chicago University Department of Education, Statistical Library.

Wright, D.J. (1986). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, Ca.

APPENDIX A1

ITEM PARAMETERS FOR 42 ITEM TEST (NO BIASED CONDITION)

No.	a	b
1	.80	-2.00
2	.80	-1.80
3	.80	-1.60
4	.80	-1.40
5	.80	-1.20
6	.80	-1.00
7	.80	-0.80
8	.80	-0.60
9	.80	-0.40
10	.80	-0.20
11	.80	.00
12	.80	.20
13	.80	.40
14	.80	.60
15	.80	.80
16	.80	1.00
17	.80	1.20
18	.80	1.40
19	.80	1.60
20	.80	1.80
21	.80	2.00
22	.80	-2.00
23	.80	-1.80
24	.80	-1.60
25	.80	-1.40
26	.80	-1.20
27	.80	-1.00
28	.80	-0.80
29	.80	-0.60
30	.80	-0.40
31	.80	-0.20
32	.80	0.00
33	.80	0.20
34	.80	0.40
35	.80	0.60
36	.80	0.80
37	.80	1.00
38	.80	1.20
39	.80	1.40
40	.80	1.60
41	.80	1.80
42	.80	2.00

APPENDIX A2

ITEM PARAMETERS FOR 42 ITEM TEST (BIASED CONDITION)

No.	a	b
1	.80	-2.00
2	.80	-1.80
3	.80	-1.60
4	.80	-1.40
5	.80	-1.20
6	.80	-1.00
7	.80	-0.14
8	.80	-0.60
9	.80	-0.40
10	.80	-0.20
11	.80	0.71
12	.80	0.20
13	.80	0.40
14	.80	0.60
15	.80	1.55
16	.80	1.00
17	.80	1.20
18	.80	1.40
19	.80	1.60
20	.80	1.80
21	.80	2.00
22	.80	-2.00
23	.80	-1.80
24	.80	-1.60
25	.80	-1.40
26	.80	-1.20
27	.80	-1.00
28	.34	-0.80
29	.80	-0.60
30	.80	-0.40
31	.80	-0.20
32	.34	0.00
33	.80	0.20
34	.80	0.40
35	.80	0.60
36	.34	0.80
37	.80	1.00
38	.80	1.20
39	.80	1.40
40	.80	1.60
41	.80	1.80
42	.80	2.00

APPENDIX B1

ITEM PARAMETERS FOR 66 ITEM TEST (NO BIAS CONDITION)

No.	a	b
1	.80	-2.00
2	.80	-1.88
3	.80	-1.75
4	.80	-1.63
5	.80	-1.50
6	.80	-1.38
7	.80	-1.25
8	.80	-1.13
9	.80	-1.00
10	.80	-0.84
11	.80	-0.75
12	.80	-0.63
13	.80	-0.50
14	.80	-0.38
15	.80	-0.25
16	.80	-0.13
17	.80	0.00
18	.80	0.13
19	.80	0.25
20	.80	0.38
21	.80	0.50
22	.80	0.63
23	.80	0.75
24	.80	0.84
25	.80	1.00
26	.80	1.13
27	.80	1.25
28	.80	1.38
29	.80	1.50
30	.80	1.63
31	.80	1.75
32	.80	1.88
33	.80	2.00
34	.80	-2.00
35	.80	-1.88
36	.80	-1.75
37	.80	-1.63
38	.80	-1.50
39	.80	-1.38
40	.80	-1.25
41	.80	-1.13
42	.80	-1.00
43	.80	-0.84
44	.80	-0.75
45	.80	-0.63
46	.80	-0.50

Appendix B1 (cont'd)

No.	a	b
47	.80	-0.38
48	.80	-0.25
49	.80	-0.13
50	.80	0.00
51	.80	0.13
52	.80	0.25
53	.80	0.38
54	.80	0.50
55	.80	0.63
56	.80	0.75
57	.80	0.84
58	.80	1.00
59	.80	1.13
60	.80	1.25
61	.80	1.38
62	.80	1.50
63	.80	1.63
64	.80	1.75
65	.80	1.88
66	.80	2.00

APPENDIX B2

ITEM PARAMETERS FOR 66 ITEM TEST (BIASED CONDITION)

No.	a	b
1	.80	-2.00
2	.80	-1.88
3	.80	-1.75
4	.80	-1.63
5	.80	-1.50
6	.80	-1.38
7	.80	-1.25
8	.80	-1.13
9	.80	-1.00
10	.80	-0.14
11	.80	-0.75
12	.80	-0.63
13	.80	-0.50
14	.80	-0.38
15	.80	-0.25
16	.80	-0.13
17	.80	0.71
18	.80	0.13
19	.80	0.25
20	.80	0.38
21	.80	0.50
22	.80	0.63
23	.80	0.75
24	.80	1.55
25	.80	1.00
26	.80	1.13
27	.80	1.25
28	.80	1.38
29	.80	1.50
30	.80	1.63
31	.80	1.75
32	.80	1.88
33	.80	2.00
34	.80	-2.00
35	.80	-1.88
36	.80	-1.75
37	.80	-1.63
38	.80	-1.50
39	.80	-1.38
40	.80	-1.25
41	.80	-1.13
42	.80	-1.00
43	.40	-0.84
44	.80	-0.75
45	.80	-0.63
46	.80	-0.50

Appendix B2 (cont'd)

No.	a	b
47	.80	-0.38
48	.80	-0.25
49	.80	-0.13
50	.40	0.00
51	.80	0.13
52	.80	0.25
53	.80	0.38
54	.80	0.50
55	.80	0.63
56	.80	0.75
57	.40	0.84
58	.80	1.00
59	.80	1.13
60	.80	1.25
61	.80	1.38
62	.80	1.50
63	.80	1.63
64	.80	1.75
65	.80	1.88
66	.80	2.00