Fall 10-25-2010

Controlling Type 1 Error Rate in Evaluating Differential Item Functioning for Four DIF Methods: Use of Three Procedures for Adjustment of Multiple Item Testing

Jihye Kim
Georgia State University

ACCEPTANCE

This dissertation, CONTROLLING TYPE I ERROR RATE IN EVALUATING DIFFERENTIAL ITEM FUNCTIONING FOR FOUR DIF METHODS: USE OF THREE PROCEDURES FOR ADJUSTMENT OF MULTIPLE ITEM TESTING, by JIHYE KIM, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree Doctor of Philosophy in the College of Education, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chair, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty. The Dean of the College of Education concurs.

_____
Chris. T. Oshima, Ph.D.
Committee Chair

_____
Yu-Sheng Hsu, Ph.D.
Committee Member

_____
William Curlette, Ph.D.
Committee Member

_____
Sheryl A. Gowen, Ph.D.
Committee Member

_____
Date

_____
Sheryl A. Gowen, Ph.D.
Chair, Department of Educational Policy Studies

_____
R.W. Kamphaus, Ph.D.
Dean and Distinguished Research Professor
College of Education

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the college of Education's director of graduate studies and research, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that nay copying form or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

_____

Jihye Kim

NOTICE TO BORROWERS

All dissertation deposited in the Georgia University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Jihye Kim
2574 Willow Grove Road
Acworth, GA 30101

The director of this dissertation is:

Dr. Chris T. Oshima
Department of Educational Policy Studies
College of Education
Georgia State University
Atlanta, GA 30303-3083

VITA

Jihye Kim

ADDRESS: 2574 Willow Grove Road
Acworth, Georgia 30101

EDUCATION:

Ph.D.	2010	Georgia State University Educational Policy Studies
M.S.	2004	Georgia State University Biostatistics
B.S.	1997	Sangji University Applied Statistics

PROFESSIONAL EXPERIENCE:

2010-Present	Doctoral Fellow Georgia State University, Atlanta, GA
2009-2005	Graduate Research Assistant Board of Regent, Atlanta, GA
2003-2005	Graduate Research Assistant Georgia State University, Atlanta, GA

PROFESSIONAL SOCIETIES AND GORGANIZATIONS:

2004-Present	National Council on Measurement in Education
2004-Present	American Educational Research Association

PRESENTATIONS AND PUBLICATIONS:

Kim, J., Oshima, T. C. (2010, July). Controlling Type I Error Rate in Evaluating Differential Item Functioning for Four DIF Methods: Use of Adjustment of Multiple Item Testing. Paper presented at the annual meeting of the Psychometric Society, Athens, GA.

Kim, J., Monsaas, J., & Paterson, P. (2010, April). The Effectiveness of an Alternative Teacher Preparation Program to Increase the Number and Diversity of Teachers in High Poverty Schools. Paper presented at the annual meeting of the American Education Research Association, Denver, CO.

ABSTRACT

CONTROLLING TYPE I ERROR RATE IN EVALUATING DIFFERENTIAL ITEM FUNCTIONING FOR FOUR DIF METHODS: USE OF THREE PROCEDURES FOR ADJUSTMENT OF MULTIPLE ITEM TESTING
by
Jihye Kim

In DIF studies, a Type I error refers to the mistake of identifying non-DIF items as DIF items, and a Type I error rate refers to the proportion of Type I errors in a simulation study. The possibility of making a Type I error in DIF studies is always present and high possibility of making such an error can weaken the validity of the assessment. Therefore, the quality of a test assessment is related to a Type I error rate and to how to control such a rate. Current DIF studies regarding a Type I error rate have found that the latter rate can be affected by several factors, such as test length, sample size, test group size, group mean difference, group standard deviation difference, and an underlying model. This study focused on another undiscovered factor that may affect a Type I error rate; the effect of multiple testing.

DIF analysis conducts multiple significance testing of items in a test, and such multiple testing may increase the possibility of making a Type I error at least once. The main goal of this dissertation was to investigate how to control a Type I error rate using adjustment procedures for multiple testing which have been widely used in applied statistics but rarely used in DIF studies.

In the simulation study, four DIF methods were performed under a total of 36 testing conditions; the methods were the Mantel-Haenszel method, the logistic regression procedure, the Differential Functioning Item and Test framework, and the Lord's chi-square test. Then the Bonferroni correction, the Holm's procedure, and the BH method were applied as an adjustment of multiple significance testing. The results of this study showed the effectiveness of three adjustment procedures in controlling a Type I error rate.

CONTROLLING TYPE I ERROR RATE IN EVALUATING DIFFERENTIAL ITEM FUNCTIONING FOR FOUR DIF METHODS: USE OF THREE PROCEDURES FOR ADJUSTMENT OF MULTIPLE ITEM TESTING

by
Jihye Kim

A Dissertation

Presented in Partial Fulfillment of Requirements for the
Degree of
Doctor of Philosophy
in
Research, Measurement, and Statistics
in
the Department of Educational Policy Studies
in
the College of Education
Georgia State University

Atlanta, GA
2010

ACKNOWLEDGEMENTS

I have so many people who encouraged and supported me. First of all, I would like to be grateful to my advisor, Dr. Chris T. Oshima, with my whole heart. I can remember when I just started the Ph.D program in the fall semester at the year 2004. I asked her to be my advisor, and she was willing to accept me. Since then it has been almost six years under her advising. She has guided me with her dedication, patience, and profound knowledge all the time. And Dr. Oshima enabled me to complete this dissertation. I could not have done it without her. Thank you so much! I also want to give my gratitude to my committee members. Thanks to Dr. Sheryl Gowen for her sincerity and encouragement! For Dr. Curlette, I appreciate his support for entire process of my study. My appreciation extended to Dr. Yu-Sheng Hsu for his guidance of my dissertation. His insightful advice was helpful for me to go right on track!

I would love to thank friends. First, I give my thanks to Vincent. He reviewed my prospectus and provided invaluable comments. It really helped! At the Board of Regent, University System of Georgia, I had been worked for several years as a graduate research assistant. Experience at the BOR was great and inspired me to look into the evaluation area. My specialized major is the Differential Item Functioning. So, the research on the teacher education program at the BOR was kind of new but invaluable experience to me. Specially, I am grateful to give appreciation to Dr. Patricia Paterson, and Dr. Judy Monsaas. For Dr. Paterson, I have learned so much about the project on the alternative teacher preparation program. And I recognize that my research and study would not have been possible without her financial assistance. Dr. Monsaas is such a wonderful person. She has been always encouraged me and kind enough to allow me to do research with her. For our research, when I seek the research topic with teacher retention data, she provided me a clear direction of research. Then we started to do research together and presented the paper at AERA. It was great for me to present at AERA for the first time. I would like to express my gratitude for her support.

Finally, I would like to thank my husband who supported me with his love, patience, and dedication. He made me finish this dissertation. I give my love to him. I also thank to my daughter, Sunny. She is a source of my energy in every single moment. She and I have been through all the way together, since I had her in the third year of my Ph.D study. Most of all, I want to give my special thank to my father. He has always inspired and encouraged me to achieve my goal of studying in U.S.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

BH	Benjamini and Hochberg Method
CDIF	Compensatory Differential Item Functioning
CMH	Cochran Mantel-Haenszel Method
DFIT	Differential Functioning of Items and Tests
DIF	Differential Item Functioning
DTF	Differential Test Functioning
ICC	Item Characteristic Curve
IPR	Item Parameter Replication
IRT	Item Response Theory
MH	Mantel-Haenszel Method
NCDIF	Non-Compensatory Differential Item Functioning
SD	Standard Deviation

CHAPTER 1

INTRODUCTION

Background and Research Questions

Differential Item Functioning (DIF) identifies items that show different outcomes in different groups (e.g., reference and focal group). In other words, DIF exists in a test when examinees of equal ability in different groups, in terms of education, ethnicity, and race, have different probabilities of correctly answering items in a test (Holland &Wainer, 1993).

The ultimate goal of a DIF study is to provide a fair opportunity for every examinee to perform successfully. A number of tests are provided for evaluating examinees' knowledge in various groups. If a test favors one group of examinees over another, the test is considered to be biased. When a test is unbiased, the score can be strong evidence of what a tester wants to assess.

In particular, a DIF study often investigates DIF items focusing on the minority groups, called the focal groups, such as a group of female or various racial/ethnic group.. It is based on the assumption that test takers who have similar knowledge (based on total test scores) should perform in similar ways on individual test questions regardless of their group membership.

Various statistical methods to detect DIF items have been developed in the testing fields of school achievement and credential examinations (Swaminathan & Rogers,

1990). Each field has different statistical criteria for detecting and evaluating DIF items (Shepard, Camilli, & Averill, 1981): for instance, an effect size measure, a standard error of the estimate, and significance testing (Mapuranga, Dorans, & Middleton, 2008). This dissertation focused on the significance testing approach, which has been an attractive research methodology for a long time, because it is easy to implement (Schmidt, 1996). In fact, most psychological and experimental papers have presented a critical value and a test statistic, which are associated with a significance test (Nickerson, 2000). In any significance testing, an error, such as a Type I error, is of interest.

A Type I error refers to the mistake of rejecting a correct null hypothesis, and the probability of committing such an error is often denoted by alpha (α). To control a Type I error, the nominal level of α is generally chosen to be small--by convention, .05. A Type I error in DIF studies refers to the mistake of wrongly identifying non-DIF items as DIF items, and a Type I error rate in DIF studies refers to the proportion of Type I errors in a simulation study. For example, if the number of Type I errors occurring in 1000 simulated items is 48 for a certain method, the Type I error rate of that method is .048. The possibility of incorrectly detecting the presence of DIF is always present. Falsely identifying DIF items can weaken the validity of the assessment. Hence, the quality of a test assessment is related to a Type I error rate and to how to control it. Additionally, an appropriate criterion of Type I error affects the quality of a test assessment.

Some research uses a criterion of a Type I error, which is based on the exact binomial distribution assuming the procedures adhere well to the nominal level of α (Nandakumar & Roussos, 2001). For example, the actual probability of a Type I error is expected to fall between .03 and .07 at the nominal level of α of .05. Some research uses

Bradley's liberal criterion (1978). "If a probability of Type I error falls within the criterion of .025 $\leq$ Probability of Type I error $\leq$ .075 at nominal α level of $.05$ and .0055 $\leq$ Type I error $\leq$ .015 at nominal α level of .01". According to Bradley (1978), the test is referred to as robust if a Type I error rate is approximately equal to the nominal α level.

Current DIF studies regarding Type I error rate have found that the latter rate can be affected by several factors, such as test length, sample size, test group size, group mean difference, standard deviation difference, distribution of difference, and an underlying IRT model that reflects person's ability/trait. Even a certain type of statistical method can affect Type I error rate. For example, the Mantel-Haenszel (Holland & Thayer, 1988) statistic follows the χ^2 distribution, which is affected by sample size.

A Type I error rate might be affected by another possible factor. DIF analysis conducts multiple significance testing of items in a test, and such multiple testing may increase the possibility of committing a Type I error at least once (Shaffer, 1995). Therefore, DIF analysis based on significance testing of every item in a test may be affected by some undiscovered factors, which could potentially spiral a type I error rate.

The main goal of this dissertation was to learn more about how to control a Type I error rate in DIF studies that are based on significance testing. Particularly, four DIF methods were considered in this dissertation: the Mantel- Haenszel (MH) method, the logistic regression procedure, the Differential Functioning Item and Test framework (DFIT), and the Lord's chi-square test. The former two methods are based on parametric IRT based methods, and the latter two methods are based on non-parametric and non IRT based methods.

For the goal of this dissertation, I investigated two distinct approaches. The first approach was to study the effects of various environmental factors: an unequal group sample size, a group mean difference, and a group standard deviation difference. The effects of these factors on the error rate have been studied inadequately in the past literature. The second approach was to study the effect of the three adjustment procedures on the error rate, which are the Bonferroni correction, the Holm's procedure, and the BH method. The second approach focused on "fishing expeditions" (Stevens, 1999). I presented below two sets of research questions:

1a. Do testing conditions affect a Type I error rate of the two methods, which are based on parametric IRT based procedures? And if so, which method performs better?

1b. Do testing conditions affect a Type I error rate of the two methods, which are based on non-parametric, non IRT based procedures? And if so, which of either method performs better?

2a. Do adjustment procedures reduce a Type I error rate of parametric IRT based methods?

2b. Do adjustment procedures reduce a Type I error rate of non-parametric, non IRT based methods?

2c. Of the Bonferroni correction, the Holm's procedure, and the BH method, which adjustment procedure work better for methods of parametric IRT based and non-parametric non IRT based both?

CHAPTER 2

LITERATURE REVIEW

Introduction to DIF methods based on statistical criteria

If an item in a test is considered biased, it violates the fundamental principle that the test should be fair to examinees. A biased item, which shows DIF, threatens the validity of test scores. There are two different types of DIF: uniform DIF and nonuniform DIF. Uniform DIF is that one group has a consistently better chance of correctly answering an item. Non-uniform DIF is that one group does not have a consistently better chance of correctly answering an item and presents differently at same ability.

For over 25 years, many statistical methods for detecting measurement bias in psychological and educational testing fields have been developed (Millsap & Everson, 1993; Swaminathan & Rogers, 1990). According to a literature review of DIF (Mapuranga et al., 2008), there are many different ways of investigating DIF by using statistical criteria. Such criteria include "the existence of an interpretable measure of the amount of DIF, the existence of a standard error estimate, and the existence of a test of significance" (p. 10). Methods of detecting DIF items based on each criterion have been being developed since the 1980s.

The first criterion, the effect size measure, is to analyze DIF items by interpreting the measure of the amount of DIF. Many non-parametric odds ratio methods are often used to measure the effect size of DIF. For example, the MH delta (Holland & Thayer, 1988), Liu-Agresti common odds ratio (Penfield & Algina, 2003), R^2-like indices

(Jodoin&Gierl, 2001; Zumbo, 1999), standardized proportion difference correct indices (Monahan, McHorney, Stump, & Perkins, 2007), are used to measure DIF magnitude.

The second criterion, the standard error of estimate, analyzes DIF items by assessing the amount of random variability associated with DIF estimates. It is a measure of the accuracy of prediction made with a regression line. Most methods of the generalized linear model, such as logistic regression, use standard error of estimates.

The third criterion, significance testing, is to make statistical inference by testing hypotheses. It is one of the most popular statistical analyses and has been used in many areas, such as psychology, social science, economics, business, and clinical studies (Minium, Clarke, & Coladarci, 1998; Royall, 1986). Many methods like the MH method, Cochran Mantel Haenszel (CMH) (Meyer, Huynh, & Seaman, 2004; Parshall & Miller, 1995), Likelihood Ratio (Thissen et al., 1988, 1993), mixture models, and the DFIT method (Raju, van der Linden, & Fleer, 1995; Oshima, Raju & Flowers, 1997; Flowers, Oshima, & Raju, 1999) use the criterion of significance testing. A classification of statistical methods based on statistical criteria was presented in Appendices A to B.

Significance testing to evaluate DIF items

Despite the recent criticisms of significance testing (Hunter, 1997; Morrison & Henkel, 2006; Rozeboom, 1960), conducting a DIF study based on significance testing is still valuable because it is simple and practical (Shepard et al., 1981).The procedure of significance testing is as follows. Sample data are collected through an observational study or an experiment, and statistical inference is done to assess claims about the population from which the sample is collected. An appropriate statistical significance test

is carried out for the given test settings: a null hypothesis, a theoretical distribution (of population and estimators), a sample size, and an a priori chosen level of α. Popular methods based on significance testing of DIF include the MH method, Cochran Mantel Haenszel methods, Lord's chi-square (Lord, 1977; McLaughlin & Drasgow, 1987), the Likelihood Ratio test, the Logistic Regression procedure (Swaminathan & Rogers, 1990; French & Miller, 1996; Rogers & Swaminathan, 1993), and Differential Functioning Item and Test (DFIT) framework.

Descriptions of four DIF methods based on significance testing

Mantel-Haenszel (MH) Method (P. W. Holland & Thayer, 1988)

The MH method is an approach of DIF detection for both dichotomous and polytomous items, by assessing the degree of association between two categorical variables (Fidalgo & Madeira, 2008). It is an extension of the traditional two way chi-square test of independence (between two variables) to the situation in which three variables are completely crossed, namely, group membership (e. g., men or women; black or white, etc), performance on the item (e. g., correct or incorrect) and any number of levels of the attribute the test is designed to measure ($2 \times 2 \times S$) contingency table, when total test score is used as the matching variable. It was introduced by Mantel and Haenszel (1959) and adapted to DIF study by Holland and Thayer (1988). This method was used at the ETS as the primary DIF detection method.

In the null DIF hypothesis, MH assumes that the odds of getting the item correct at a given level of the matching variable is the same in both the focal group and the reference group, across all score levels of the matching variable.

$$H_0 : \frac{R_{rs}/W_{rs}}{R_{fs}/W_{fs}} = \alpha_s = 1 , \qquad s = 1, \cdots, S \tag{1}$$

where R_{fs} is the number of people in the focal group at score level *s* who answered the item correctly. W_{fs} is the number of people in the focal group at score level *s* who answered the item incorrectly. R_{rs} is the number of people in the reference group at score level *s* who answered the item correctly. W_{rs} is the number of people in the reference group at score level s who answered the item incorrectly. N_{rs} is the total number of R_{rs} and W_{rs}. N_{fs} is the total number of R_{fs} and W_{fs}. N_{ts} is the total number of R_{rs}, W_{rs}, R_{fs}, and W_{fs}.

If $\alpha_s > 1$, the reference group has an advantage on the item; if $\alpha_s < 1$, the advantage lies with the focal group. There is a chi-square test associated with the MH method, namely a test of the null hypothesis, $H_0 : \alpha_s = 1$,

$$\text{MH}{-}\chi^2 = \frac{[|\sum_s R_{rs} - \sum_s E(R_{rs})| - .5]^2}{\sum_s Var(R_{rs})} , \tag{2}$$

where $E(R_{rs}) = E(R_{rs} \mid \alpha = 1) = \frac{N_{rs}R_{ts}}{N_{ts}}$, and

$Var(R_{rs}) = Var(R_{rs} \mid \alpha = 1) = \left[N_{rs}R_{ts}N_{fs}W_{ts}\right] / \left[{N_{ts}}^2(N_{ts} - 1)\right]$,

where the $-.5$ in the equation for MH$-\chi^2$ serves as a continuity correction to improve the accuracy of the chi-square statistic. MH$-\chi^2$ is distributed approximately as a chi-square with one degree of freedom.

As a modification of the Mantel-Haenszel test, there is the Cochran Mantel-Haenszel Test. While the Mantel-Haenszel test measures the strength of association by estimating the common odds ratio, the Cochran Mantel-Haenszel statistic assumes a

common odds ratio and tests the null hypothesis that two variables are conditionally independent in each stratum, assuming that there is no three-way interaction (Agresti, 1996).

Logistic Regression Procedure

Logistic Regression is the procedure that is used widely in statistical literature (Hariharan Swaminathan & Rogers, 1990). The method uses a model that links a categorical outcome (e.g., dichotomous) with one or more predictor variables, which can be either continuous or categorical.

$$p = \frac{e^z}{(1+e^z)} \tag{3}$$

$$z = ln\left[{p_i}/{(1-p_i)}\right] = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)\ , \tag{4}$$

where

θ = ability level (total score; the observed trait level of an examinee)

g = grouping valuable (for instance, dummy coded as 1=reference, 2=focal)

θg = the product of the two independent variables

τ_2 = corresponds to the group difference in performance on the item

τ_3 = corresponds to the interaction between group and trait level

An item shows uniform DIF if $\tau_2 \neq 0$ and $\tau_3 = 0$. An item shows non-uniform DIF if $\tau_3 \neq 0$ (whether or not $\tau_2 = 0$). The hypothesis of interest is $\tau_2 = \tau_3 = 0$. Swaminathan and Rogers (1990) shows a natural hierarchy of entering variables into the logistic model (Zumbo, 1999). There are three steps for entering variables, listed below (1999, p. 26).

1. One first enters the conditioning variable (the total score)
2. The group variable is entered
3. The interaction term is entered into the equation.

Through these three steps, Swaminathan and Rogers (1990) computed the χ^2 statistic with 2 degrees of freedom. When the value of the statistic exceeds the critical value of $\chi^2_{\alpha,2}$, the hypothesis that no DIF exists is rejected. In the logistic regression equation, DIF is measured by the simultaneous test of uniform and non-uniform DIF.

Non-compensatory DIF (NCDIF) Index in DFIT

The DFIT method (Raju, Linden, & Fleer, 1995) has three indices, which are non-compensatory DIF (NCDIF), compensatory DIF (CDIF), and a differential test function (DTF) index. The NCDIF index begins with the assumption that all other items have no DIF. The concept of CDIF is related to DTF. The sum of CDIF values are the value of DTF that enables a researcher to examine the net effect of deleting items from the test (Oshima, Raju, & Nanda, 2006; Raju et al., 1995).

DFIT has several benefits that enable it to assess differential item functioning not only at the item level but also at the test level. It can be used for both dichotomous and polytomous scoring schemes and handles both uni-dimensional and multidimensional models. Among three indices of the DFIT method, this dissertation focused on the NCDIF index, which is based on the chi-square significance testing. NCDIF is defined as

$$NCDIF_i = {\sigma_{d_i}}^2 + {\mu_{d_i}}^2, \quad (5)$$

which assumes that if all items in the test other than item *i* are completely unbiased, then it must be true that $d_j = 0$ at all $i \neq j$. d_i is defined as the difference in item probabilities for item *i*. If the item parameters for item *i* are equal for both the focal and reference groups, then it assumes that there is no DIF(NCDIF = 0). The chi-square test for NCDIF is

$$\chi_{N_F}{}^2 = \frac{N_F(NCDIF)}{\sigma_{d_i}{}^2}, \tag{6}$$

With N_F degrees of freedom, which is the sample size of focal group, given d_i is normally distributed with a finite variance. This chi-square significance testing for NCDIF had been used until the Item Parameter Replication (IPR) method (Oshima et al., 2006) was proposed.

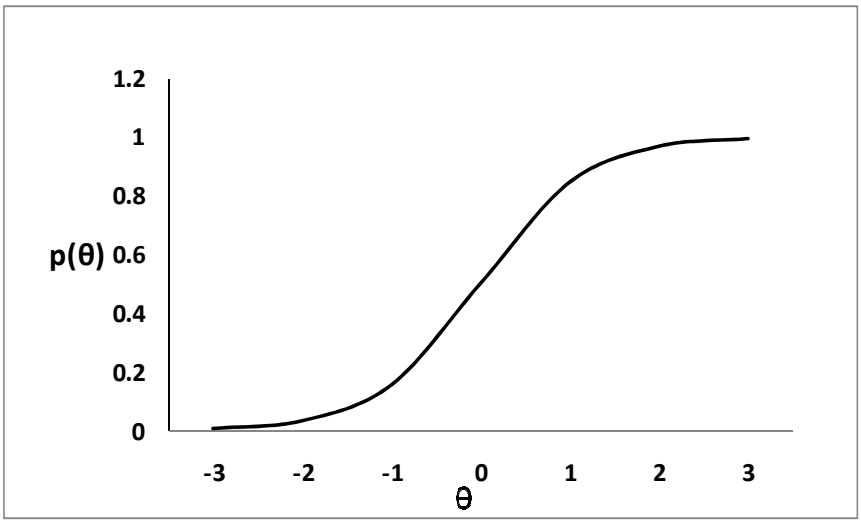
As a new version of the DFIT method, Oshima et al.(2006) proposed new cutoff values for each item by $(1 - \alpha)$ percentile rank score from a frequency distribution of NCDIF values under the no DIF conditions in the DFIT framework (APPENDIX C). The fixed cutoff value of NCDIF index for an item, *i*, (Raju, van der Linden, & Fleer, 1995) was defined as .006 in the old version of DFIT for dichotomous item analysis (Fleer, 1993). In order to improve this procedure for assessing the statistical significance of the NCDIF index, the new version of the DFIT method (IPR method) developed cutoff values ranged from .003 to .15, which conditions are "with a higher cutoff value for a smaller sample size and a higher value for an IRT model with more parameters" (p. 2). As a result, it provides fitted cutoff scores to a particular data set and reduces time consuming repeated calibrations of item parameters. This new procedure was shown as an effective way to detect DIF items.

Lord's chi-square Test

Lord's chi-square is based on an examination of the differences in the variance-covariance matrix of the difficulty and discrimination parameters. It calculates the differences in the areas between the curves for two groups (Hambleton, H Swaminathan, & Rogers, 1991). The curve means a graphical expression which represents the performance of an item in a test. It is described by the person ability parameter (θ) and the item parameters (*a, b*, and *c* parameter) as shown below.

Figure 1

The item characteristic curve for a dichotomous item



Lord's chi-square statistic is given by

$$\chi_i^2 = \left(a_{diff} b_{diff} c_{diff}\right)' \sum^{-1} \left(a_{diff} b_{diff} c_{diff}\right), \quad (7)$$

where

$\sum^{-1}$ is the inverse variance-covariance matrix for the differences in item parameter estimates

$a_{diff} = a_{reference\ group} - a_{focal\ group}$ (a is the discrimination parameter)

$b_{diff} = b_{reference\ group} - b_{focal\ group}$ (b is the difficulty parameter)

$c_{diff} = c_{reference\ group} - c_{focal\ group}$ (c is the pseudo guessing parameter)

The distribution of Lord's chi-square should be close to the chi-square distribution with two degrees of freedom at 2PL model and that with three degrees of freedom at 3PL model (Lord, 1980). Lord's chi-square test is more useful with 2PL and 3PL models than with the 1PL model, since examining the difference between groups by using only a one parameter model may provide inaccurate or insufficient results of DIF.

Type I error rate on DIF

In significance testing on DIF, the effective control of the Type I error rate has been of interest. Many research findings have addressed the fact that the Type I error rate is influenced by various conditions, and researchers have tried to control the Type I error rate by setting factors differently. For instance, Uttaro and Millsap (1994) showed that the MH method exhibited different Type I error rates for different test lengths: The result for the 20 items test showed more inflation of Type I error rates than one for the 40 items. Finch (2005) showed that the performance of the Simultaneous item bias test (SIBTEST) is better in the two-parameter model than three -parameter model. Finch and French (2007) showed that the performance of the Likelihood Ratio (IRTLR) test was low if the sample size is small. The MH method performed poorly in the non- uniform DIF condition (Hambleton & Rogers, 1989).

The adjustment procedure in multiple significance testing can be an effective way to control a Type I error rate. Although the probability of making a Type I error for each significance test is set at some nominal α level, the overall α (the family-wise Type I error rate, α_{FW}) across the entire set of significance tests can be considerably higher. The family-wise Type I error rate is the probability of making at least one Type I error across multiple significance tests. It is calculated as $\alpha_{FW} = 1 - (1 - \alpha)^c$, where c is the number of tests.

As said earlier, one issue in multiple significance testing is a higher chance of committing a Type I error of identifying non-DIF items as DIF items. Common approaches to control such an issue include the Bonferroni correction (Bonferroni, 1936), the Holm's procedure, and the Benjamini and Hochberg False Discovery Rate method (Benjamini & Hochberg, 1995).

While the Bonferroni correction and the Holm's procedure seek to control family-wise Type I error rate, Benjamini and Hochberg (1995) controls expected false positive discovery rates (FDR) by defining a sequential p-value procedure. The Benjamini and Hochberg (BH) method for p-values $p_1, \cdots, p_m$ follows as:

1. Rank order p-value of each item form the largest to the smallest
2. Remain the largest p-value
3. Multiply the second largest p-value by the total number of items in test and then divide by its rank (corrected p-value = p-value*(n/(n-1)), where *n* is the total number of items). If the corrected p-value < .05, it is significant.

4. Multiply the third largest *p-value* and divide by its rank as in Step 3 (corrected p-value = p-value*(n/(n-2)). If the corrected p-value < .05, it is significant.

continues until the smallest p-value is corrected.

The BH method has been adopted in the study of DIF (Steinberg, 2001; Thissen, Steinberg, & Kuang, 2002; Williams, Jones, & Tukey, 1999). Williams, Jones, and Tukey (1999) compared the BH method with the Bonferroni correction and showed the BH method performed better. Steinberg (2001) also used the BH method with the likelihood ratio procedure for the evaluation of DIF.

The Bonferroni correction is an adjustment to control family-wise Type I error rate in multiple comparison procedures. If there are *k* hypotheses to be tested, each test should be conducted at significance level of α/k, where *k* is the number of hypotheses (Holland & Cohenhaver, 1988).

The Bonferroni correction is a rough approximation, so it is conservative. Perneger (1998) pointed out the problem of the Bonferroni adjustment. The main weakness is that the Type I errors cannot be reduced without inflating Type II errors, which is the probability of accepting the false negative, which is the mistake of failing to reject a null hypothesis.

Holm (1979) presented an improved procedure that is more powerful than the Bonferroni method (B. S. Holland & Copenhaver, 1987). The Holm's procedure is very similar to the Bonferroni, but it is known to be less conservative because the Holm's procedure is less corrective as the p-value increases, and is based on the ordered p-values

of the individual tests. The ordered p-value methods are strong for controlling a Type I error rate, when the test statistics are independent (Shaffer, 1995).

The Holm's procedure has two steps (Shaffer, 1995). The first step of the procedure is to order the *k* numbers of p values from the smallest to the largest and denote the ordered $\{p_i\}$ by $p_{(1)} \leq \cdots \leq p_{(k)}$. Let $H_{(1)}, \cdots, H_{(k)}$ be the corresponding hypotheses. Suppose i^* is the smallest integer from 1 to *k* such that

$$p_{(i^*)} > \alpha/(k - i^* + 1) \tag{8}$$

Then the Holm's procedure rejects $H_{(1)}, \cdots, H_{(i^*-1)}$ and retains $H_{(i^*)}, \cdots, H_{(k)}$. If $p_{(i^*)}$ is greater than $\alpha/(k - i^* + 1)$ for no integer i^*, then all *k* hypotheses are rejected. Holm (1979) proves that this procedure guarantees that there is at most α chance of rejecting at least one of the true hypotheses.

There are literature reviews on applying the Holm's procedure in the areas of clinical trials and biology. For example, Soulakova (2009) used the Holm's procedure to find a problem of identifying all effective and superior drug combinations. There is also research on applying the Holm's procedure to psychometrics for evaluating appropriate curriculum based measurement of reading outcomes (Betts, Pickart & Heistad, 2009), but none is found for the study of DIF so far.

Regarding the purpose of the dissertation, a literature review had been an account of what has been studied on topics that were related with DIF studies. Relative articles had been reviewed focusing on what methods were effective in a particular testing condition, how DIF methods control Type I error rate, what had been, what adjustment

procedures had been used to control the multiple significant testing issues in DIF studies, and how DIF methods detect DIF items in terms of different criteria (until 2008). The list of DIF methods by different criteria of evaluating DIF items is provided in Appendix A.

CHAPTER 3

METHOD

Study Design

The research design consisted of two phases. The first phase conducted significance testing of the four DIF methods; the performance of each DIF method was then compared across all conditions. In order to answer the research questions, significance testing of each DIF method was performed with a total of 16 conditions: 4 conditions of sample sizes × 2 conditions of test lengths × 2 conditions of group difference (group mean difference (impact) and group standard deviation (SD) difference). The second phase applied the Bonferroni correction, the Holm's procedure, and the BH method to the significance testing of each DIF method. A SAS program was used to conduct the simulation studies and DIF analysis.

Conditions of Study

Data Generation

Using SAS program, the dichotomous scored data were generated for the reference and the focal groups with a three-parameter IRT model. The ability of test examinee was assumed to follow the standard normal distribution. First, the probability of a correct response to an item was calculated based on pre-specified item parameters from Oshima et al. (2006); the basis probability was generated at random from the uniform distribution. Next, the calculated probability and a basis probability were

compared. If the basis probability was less than the calculated probability, the simulated item response was scored as correct (1); otherwise, it was scored as incorrect (0).

Sample Size and Sample Size Ratio

Sample size is a core factor for detecting DIF items accurately. Although a small sample size could cause a poor estimation, resulting in true DIF items not being detected well, a large sample size could result in precise detection of true DIF items, although the possibility exists that items with no or a very little DIF will be detected as if they are true DIF items.

In most DIF research, the range of sample sizes is between 500 and 5,000 for both equal and unequal sample sizes. This research chose total sample sizes of 1,000 and 2,000, etc. The ratio of the sample sizes between the focal group and the reference group was also considered for non-parametric methods as existing research on DIF with unequal sample sizes showed a greater tendency to detect flagged DIF items than one with equal sample sizes (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005). However, the parametric method had one condition of the sample size (2,000) and an equal ratio of sample sizes.

For the non-parametric method (Table 1), when the sample size was 1,000 and the group sizes were the same, the size of each group was 500; when the group sizes were different, the size of the reference group was 700 and that of the focal group was 300. When the sample size was 2,000 and the group sizes were the same, the size of each group was 1,000; when the group sizes were different, the size of the reference group was 1,500 and that of the focal group was 500.

Table 1

Sample size and sample size ratio of non- parametric methods

Group	Sample size	
	Focal	Reference
Equal sample sizes of two groups	500	500
	1000	1000
Unequal sample sizes of two groups	300	700
	500	1500

Test Length

This study chose test lengths of 20 items and 40 items. To date, much of the DIF research has been conducted using a test length of between 20 and 40 items because common assessments are constructed with fewer than 40 items. Raju et al. (1995) selected 40 items in their simulation study. Roussos and Stout (1996) conducted their simulation study with 25 items. In particular, 40 items has been chosen in many studies (Jodoin & Gierl, 2001; Narayanan & H Swaminathan, 1994; Rogers & H Swaminathan, 1993).

Percent of DIF Level

The setting of item parameters for data generation was identical to those in a study by Raju et al. (1995) and Oshima et al. (2006). The condition of 10% DIF level was investigated: All *a*, *b* and *c* item parameters were set to be the same for both reference and focal groups, except that two items (Item 3 and 8) in a 20 items test and four items (Item 5, 10, 15, and 20) in a 40 items test had different *a* and *b* item parameters in the focal group as shown at Table 2.

Table 2

Item parameters for the 40 items test and the 20 items test

Item		Reference		Focal (10%)	
40	20	*a*	*b*	*a*	*B*
1	1	.55	0		
2		.55	0		
3	2	.73	-1.04		
4		.73	-1.04		
5	3	.73	0	.73	1
6		.73	0		
7	4	.73	0		
8		.73	0		
9	5	.73	1.04		
10		.73	1.04	.73	1.54
11	6	1	-1.96		
12		1	-1.96		
13	7	1	-1.04		
14		1	-1.04		
15	8	1	-1.04	1	-.04
16		1	-1.04		
17	9	1	0		
18		1	0		
19	10	1	0		
20		1	0	1	.5
21	11	1	0		
22		1	0		
23	12	1	0		
24		1	0		
25	13	1	1.04		
26		1	1.04		
27	14	1	1.04		
28		1	1.04		
29	15	1	1.96		
30		1	1.96		
31	16	1.36	-1.04		
32		1.36	-1.04		
33	17	1.36	0		
34		1.36	0		
35	18	1.36	0		
36		1.36	0		
37	19	1.36	1.04		
38		1.36	1.04		
39	20	1.8	0		
40		1.8	0		

*C parameter is fixed with .20 for both reference and focal groups.

Group Difference

Two different conditions were set up. The first condition focused on the mean difference between groups: the same mean between focal and reference groups versus different means between the groups. The first component of the group means difference was implemented by assuming that both reference and focal groups follow standard normal distribution. The second component was implemented by assigning a lower mean (by .2) to the focal group than to the reference group.

The second condition focused on the standard deviation (SD) difference between the groups: no difference of SD between focal and reference groups versus different SD between the groups. The first component of the group SD difference was implemented by assuming that both reference and focal groups follow the standard normal distribution. The second component was implemented by assigning a lower SD (by .2) to the focal group than to the reference group (Penny & Johnson, 1999).

Replications

In order to ensure stable results, an appropriate number of replications were needed. This study included 100 replications, because 100 times was the common replication according to the publications of the National Council on Measurement in Education (NCME) in 2009.

Statistical DIF Methods for Simulation Studies

Four DIF methods were selected for this study: the MH method, the logistic regression procedure, the DFIT method, and the Lord's chi-square test. As an adjustment of multiple significance testing, the Bonferroni correction, the Holm's procedure, and the BH method were applied.

DIF detection procedure for non-parametric methods

For each generated data set, the MH method and the logistic regression procedure were used for DIF detection. Prior to DIF detection, ability matching for reference and focal groups was performed by calculating the total scores (Zwick, J. R. Donoghue, & Grima, 1993). Based on the total calculated score, matching ability was performed. This study chose a form of thin matching used by Donoghue and Allen (1993). Thin matching is based on using the total scores as the matching variable. The matching variable created eight intervals. The total scores were then categorized into corresponding intervals in order to match groups into equal intervals. After matching was completed, the statistics of both the MH method and the logistic regression procedure were calculated for each of replicated datasets.

DIF detection procedure for parametric methods

Several steps were needed in parametric DIF analyses. This study conducted two stage-linking procedures. First, the generated data sets were calibrated using BILOG-MG3 to obtain item parameter estimates. Second, item parameter estimates for reference and focal groups were put on the common scale by determining linking coefficients through the mean and sigma method. Third, for the DFIT method, the DIFCUT program

was used to determine cutoff scores and NCDIF values. The statistics of Lord's chi-square was also calculated using the Lord's chi-square test. Once DIF items were detected, the second linking procedure was done by calculating linking coefficients again with only non-DIF items. Using the new linking coefficients, refined statistics from the DFIT method and Lord's chi-square test were obtained.

CHAPTER 4

RESULTS

This study had two purposes. The first was an endeavor to investigate the degree of the type I error rate with various factors. The second was to apply adjustment approaches--specifically the Bonferroni correction, the Holm's procedure, and the BH method--to a case of multiple significance tests on a DIF study. A total of 36 conditions were simulated. In each condition, 100 replications were performed for each of the four DIF methods, producing a total of 14,400 simulated data sets.

Although individual tests of each item were conducted at a Type I error rate of .05, the overall Type I error rate and degree of power were questioned when multiple items were tested concurrently (Hoffman & Recknor & Lee, 2008). The importance of understanding the severity of an inflated Type I error rate has been discussed in other studies (Lin &Rahman, 1998). In order to investigate the inflation of a Type I error rate, Bradley's (1978) liberal robustness criterion range of .025 to .075 was used. If a Type I error rate for each DIF method was within this range, the Type I error rate was considered well-controlled.

The test-wide Type I error rate for this study was calculated as follows. First, for each replication, the occurrences of false positives out of all non-DIF items were counted. Then, the proportion of these counts was calculated per replication, focusing on the practical point of view how many items were falsely identified as DIF items in each test set. A Type I error rate is the average of these proportions. For example, Table 3 presents

the false positives of the Lord's chi-square test in a 40 items test; there are 4 DIF items (Item 5, 10, 15, and 20) and 36 non-DIF items. The number of false positives is shown at the bottom on the table: six in the 1st replication, four in the 2nd, and so on. The proportion of false positives per replication was calculated out of 36 non-DIF items. That is, the proportions of false positives were 6/36 in the 1st replication, 4/36 in the second, and so on. The average of these proportions was .13; that is, the average of the test-wise Type I error rate was .13. Therefore, on average, 13% of non-DIF items were falsely identified as DIF items.

$$\text{Type I error rate} = \frac{1}{100}(6/36 + 4/36 + \cdots + 3/36) = .13$$

Table 3

False positives and true positives for each item of Lord's chi-square test in a 40 items test

	Replication for Lord's chi-square test								
Item	1st	2nd	3rd	4th	5th	……	98th	99th	100th
1	0	0	0	0	0	……	0	0	0
2	0	0	0	0	0	……	0	0	0
3	0	0	0	0	0	……	0	0	0
4	0	0	0	0	0	……	0	0	0
5	1	1	1	1	1	……	1	1	1
6	0	1	0	0	0	……	0	0	0
7	0	0	0	1	0	……	0	0	0
8	0	0	0	0	0	……	0	0	0
9	0	0	0	0	0	……	0	0	1
10	0	0	0	1	1	……	0	1	0
11	0	0	0	0	0	……	0	0	0
12	0	0	0	0	0	……	0	0	0
13	0	0	0	0	0	……	0	0	0
14	0	0	0	0	0	……	0	0	0
15	1	1	1	1	1	……	1	1	1
16	0	0	0	1	0	……	0	0	0
17	0	0	0	0	0	……	0	0	0
18	0	0	0	0	0	……	0	0	0
19	0	0	0	0	0	……	0	0	0
20	1	1	1	1	1	……	1	1	1
21	0	0	0	0	0	……	0	0	0
22	0	0	0	0	0	……	0	0	0
23	0	0	1	0	0	……	0	0	0
24	1	0	0	0	0	……	0	0	0
25	1	1	1	1	1	……	1	1	1
26	0	0	0	0	0	……	0	0	0
27	0	0	0	0	0	……	0	0	0
28	0	0	0	0	0	……	0	0	0
29	0	0	0	0	0	……	0	0	0
30	0	0	0	1	0	……	0	0	0
31	0	0	0	0	0	……	0	0	0
32	0	0	0	0	0	……	0	0	0
33	0	0	1	0	0	……	0	0	0
34	1	0	0	0	0	……	0	1	0
35	1	1	1	1	1	……	1	1	1
36	0	0	0	0	0	……	0	1	0
37	1	1	1	0	0	……	0	0	0
38	0	0	0	0	0	……	0	0	0
39	1	0	0	0	0	……	0	0	0
40	0	0	0	1	0	……	1	0	0
# of FP	6	4	5	6	2	……	3	4	3
# of TP (Large DIF magnitude)	2	2	2	2	2	……	2	2	2
# of TP (Medium DIF magnitude)	1	1	1	2	2	……	1	2	1

** FP: False Positives, TP: True Positives * Highlighted rows indicated the DIF Items*

The power rate in this study was also of interest. The investigation examined the trend of power rate with two types of DIF magnitude that reflects the difference on item parameters, since a power rate is affected by factors, such as the difference between two groups, sample size, etc. In this study, the difficulty (*b*) parameter value of each DIF item varied for the reference and focal groups, and I focused on the two magnitudes (large and medium) of difference between the two groups. In the simulation, two items (Items 3 and 8) in the 20 items test and four items (Item 5, 10, 15, and 20) in the 40 items test were set up as DIF items. The item difficulty (*b*) parameter values of these DIF items are shown in Table 4. For example, difficulty (*b*) parameters of Item 3 in a 20 items test are 0 and 1 for the reference and focal groups, respectively; their difference is 1. In the 40 items test, differences of *b* parameters for Items 5 and 15 were all one, which was denoted as the large DIF magnitude; differences of *b* parameters for Items 10 and 20 were 0.5, which was denoted as the medium DIF magnitude. Therefore, the investigation of separated power rates by different DIF amount was needed in the 40 items test.

Table 4
Difference of b parameter for a 20 items test and a 40 items test

Test Length	DIF items	Reference	Focal	Difference of *b* Parameters	DIF Magnitude
20	3	0.00	1.00	1	Large
	8	-1.04	-0.04	1	Large
40	5	0.00	1.00	1	Large
	10	1.04	1.54	0.5	Medium
	15	-1.04	-0.04	1	Large
	20	0.00	0.50	0.5	Medium

The power rate was calculated in the similar way as the Type I error rate was calculated. First, the proportion of true positives out of all DIF items with a specific DIF magnitude was calculated for each replication. The power rate with a specific DIF magnitude is the average of these proportions.

The highlighted rows in the Table 3 indicate DIF items. To calculate the power rate with the large DIF magnitude, true positives out of two items (Items 5 and 15) in each replication were counted first: two in the 1st replication, two in the 2nd, and so on. Then, the proportion of true positives was calculated for each replication. The power rate was the average of these proportions as shown below:

$$\text{Power rate (large DIF magnitude)} = \frac{1}{100}(2/2 + 2/2 + \cdots + 2/2) = .99$$

Similarly, for the power rates with the medium DIF magnitude, true positives out of two items (Items 10 and 20) were counted first: 1 in the 1st replication, 1 in the 2nd, and so on. Then, the power rate was calculated as shown below:

$$\text{Power rate (medium DIF magnitude)} = \frac{1}{100}(1/2 + 1/2 + \cdots + 1/2) = .55$$

All simulated false positives and true positives are presented in a tabular format in Appendices D-G.

Phase I

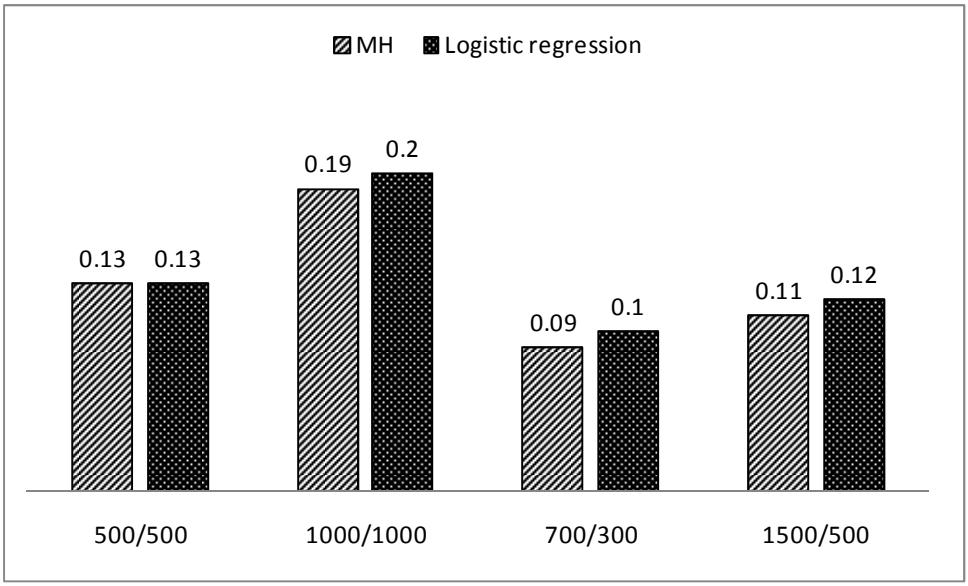
Research questions 1a and 1b listed previously addressed the effectiveness of control for the Type I error rate for the four DIF methods under several testing conditions. This study was also interested in comparing the Type I error rates of two non-parametric methods (the MH method and the logistic regression procedure) with those of two parametric methods (the DFIT method and the Lord's chi-square test), to see which DIF method performed better under specific conditions.

Comparing the Performance of Four DIF methods

The results showed that the two non-parametric methods performed similarly under all testing conditions considered in this study. For example, Figure 2 illustrates the similarity between the MH method and the logistic regression procedure when Type I error rates in the 20 items test were examined for different sample sizes. When the sample sizes for reference and focal groups were both 500 (i.e., 500/500), the Type I error rates for the MH method and logistic regression were both .13. For the sample size of 1000/1000, the Type I error rates were .19 and .20 for the MH method and the logistic regression, respectively. Similar interpretations apply to the sample sizes of 700/300 and 1500/500.

Figure 2

Type I error rate for non-parametric methods based on sample size with 20 items test



The two parametric methods, according to the results, also performed similarly under all testing conditions considered in this paper except for the condition of different test length. As shown in Figure 3, for the 20 items test, the Type I error rate of the DFIT method was somewhat higher than that of the Lord's chi-square test (.06 vs. .03). The graphical comparison across the four DIF methods when the sample size is 1000/1000 is presented in Figure 4.

Figure 3

Type I error rate for parametric methods based on sample size 1000/1000

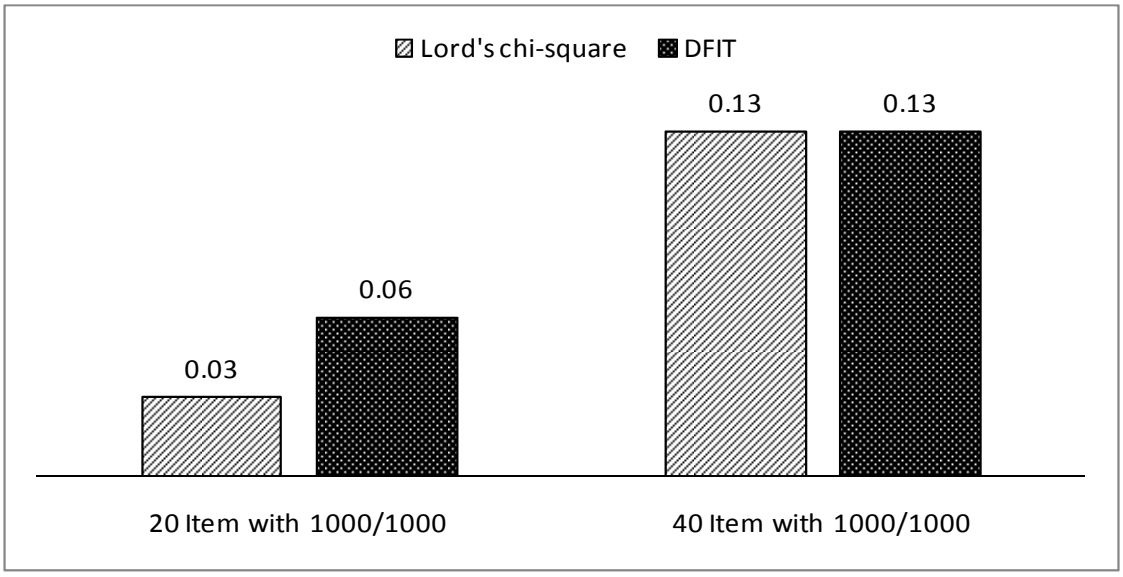
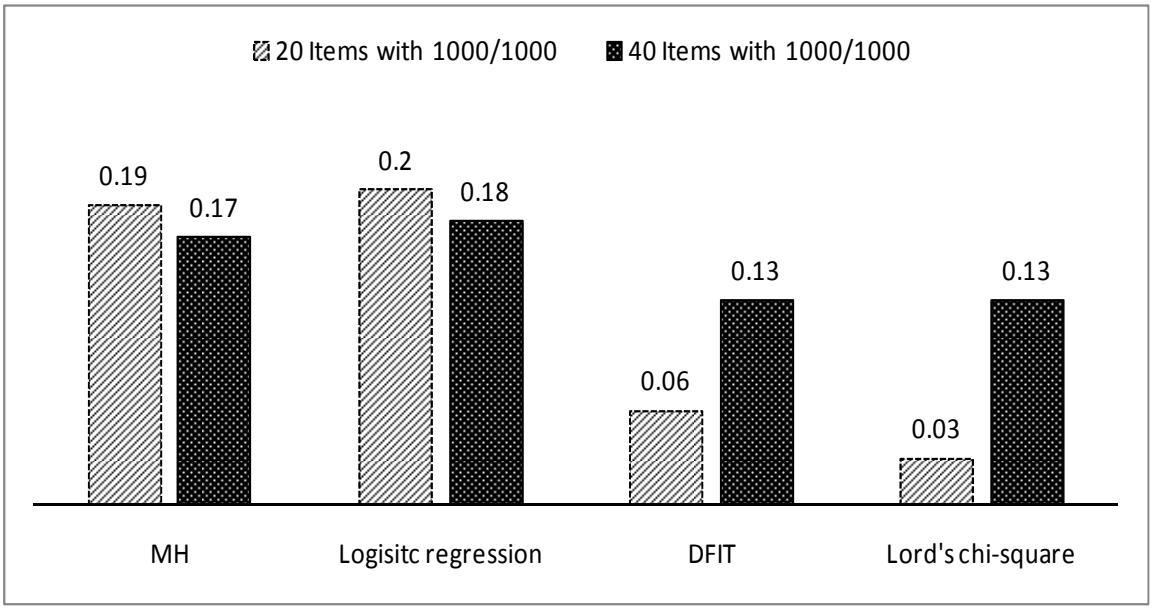


Figure 4

Type I error rates among four DIF methods with 1000/1000 sample size.



In sum, the four DIF methods--whether the method was non-parametric or parametric--performed similarly for the 40 items test. However, with the 20 items test, the Type I error rates of the non-parametric methods were higher than those of the parametric methods (Figure 4). With further investigation, the effects of testing conditions on Type I error rates of the four DIF methods were explained separately for each testing condition.

The Effect of sample size conditions, sample size ratio, and test length

As previous research suggests (Bolt, 2000; Finch, 2005; Gotzmann & Boughton, 2004), both sample size and sample size ratio affected the Type I error rates for the DIF methods. The MH method and the logistic regression procedure showed higher Type I error rates for an equal ratio of sample sizes (500/500, 1000/1000) than for those of unequal sample size ratios (700/300, 1500/500).

Table 5 shows Type I error rates and power rates of the MH method and the logistic regression for the 20 items and 40 items tests with various sample sizes. For example, for the 20 items test, the Type I error rate of the MH method was .13 when a sample size was 500/500 for reference and focal groups. This rate was higher than the Type I error rate of the same method for a sample size of 700/300 (.09). Additionally, the Type I error rate of the MH method for the 20 items test in Table 5 was .19 for a sample size of 1000/1000, which was higher than that the rate of the same method for a sample size of 1500/500 (.11). For the logistic regression procedure for the 20 items test, the Type I error rate for a sample size of 500/500 was .13, which was higher than that (.10)

for a sample size of 700/300. The Type I error rate with a sample size of 1000/1000 was .20, which was also higher than that (.12) for a sample size of 1500/500.

Table 5 also presents the power rate of both the MH method and logistic regression procedure. In the 20 items test, the power rates were high for both non-parametric methods since the DIF magnitudes of two both DIF items were large. In the 40 items test, the power rates with the large DIF magnitude were high across all conditions, but the power rates with medium DIF magnitude were very low.

Table 5

Type I error rate and power rate of non-parametric methods by test length and sample size

Test Length	Sample size	Type I error Rate		Power Rate (with DIF magnitude)	
		MH method	Logistic Regression	MH method	Logistic Regression
20	500/500	.13	.13	1.00 (1)	1.00 (1)
	1000/1000	.19	.20	1.00 (1)	1.00 (1)
	700/300	.09	.10	.98 (1)	1.00 (1)
	1500/500	.11	.12	1.00 (1)	1.00 (1)
40	500/500	.12	.13	1.00 (1)	1.00 (1)
				.28 (.5)	.30 (.5)
	1000/1000	.17	.18	1.00 (1)	1.00 (1)
				.48 (.5)	.49 (.5)
	700/300	.12	.13	1.00 (1)	.97 (1)
				.24 (.5)	.21 (.5)
	1500/500	.09	.11	.95 (1)	1.00 (1)
				.23 (.5)	.28 (.5)

* (1: large) and (.5: medium) indicate DIF magnitude (the *b* item difficulty parameter difference between reference and focal group)

The parametric methods--the DFIT method and the Lord's chi-square test were examined under only one sample size condition (1000/1000) since the DFIT method was recommended for conditions of equal ratio and a sample size greater than 1000 similar to SIBTEST (Gierl et al., 2004). The results are shown in Table 6. For the sample size of 1000/1000, the Type I error rates of the DFIT method and the Lord's chi-square test for the 40 items test were .13 and .13, respectively; the Type I error rates for a 20 items test for the two methods were .06 and .03.

The power rates of the parametric methods also showed similar pattern as the non-parametric methods. The power rates were consistently high with the 20 items test. The power rates with the 40 items test varied for different DIF magnitudes. When the DIF magnitude was large, the power rates were high (above .99 across all conditions). The power rates was low (.49 for DFIT method and .55 for the Lord's chi-square test) with medium DIF magnitude.

Table 6

Type I error rate and power for parametric methods by test length and sample size

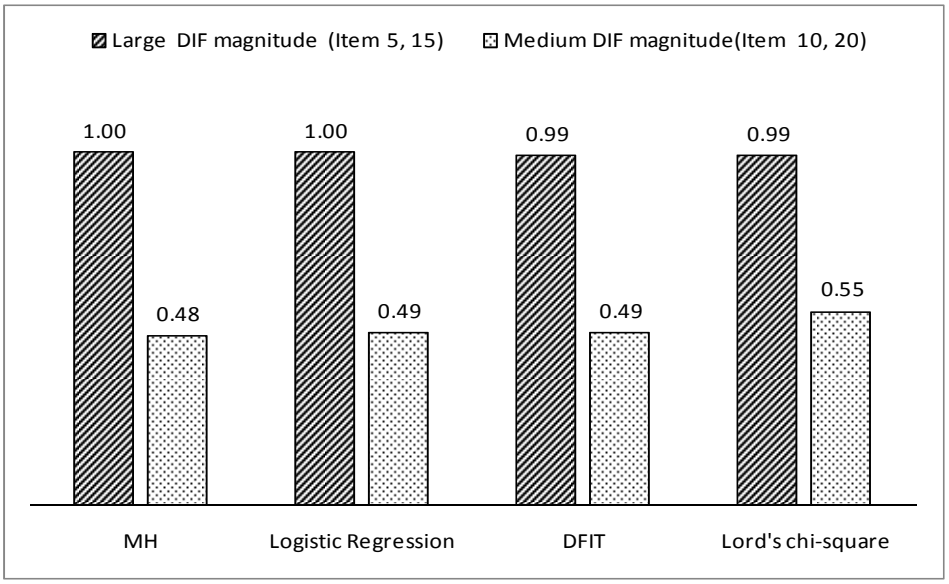
Test Length	Sample Size	Type I error Rate		Power Rate (with DIF magnitude)	
		DFIT	Lord's chi-square	DFIT	Lord's chi-square
20	1000/1000	.06	.03	1.00 (1)	1.00 (1)
40	1000/1000	.13	.13	.99 (1) .49 (.5)	.99 (1) .55 (.5)

* (1: large) and (.5: medium) indicate DIF magnitude (the *b* item difficulty parameter difference between reference and focal group)

In summary, Figure 5 shows the overall patterns of the power rate for four DIF methods for different DIF magnitude-large or medium. It shows a very strong and consistent pattern. DIF items were detected with (almost) perfect accuracy with the large DIF magnitude.

Figure 5

The power rate of four DIF methods with 1000/1000 by different DIF magnitude



The Effect of Three Conditions for Group Difference

The performances of the four methods were first examined assuming the same abilities between reference and focal groups—no group difference. The next analysis addressed the effect of the assumption that the abilities of two groups had different mean/standard deviation on the Type I error rates and power rates of the four methods. As mentioned previously, two possible group difference conditions—group mean difference and group SD difference—were set up. The analysis compared the

performances of the methods under each group difference condition with the performances under no group difference condition. The results indicated that group SD difference, but not group mean difference, highly affected the Type I error rates of three methods: the MH method, the logistic regression procedure, and the Lord's chi-square test. The DIFT method was not affected much by either group difference conditions.

The results showed that the Type I error rates of the MH method under the group mean difference condition were nearly the same as the rates under the no group difference condition (Table 7). Similarly the power rates under the two conditions were also very similar (Table 8). These results coincided with previous research: the trivial effect of group mean difference (Sheppard, Han, Colarelli, Dai, & King, 2006).

The Type I error rates of the MH method under the group SD difference condition, however, were different from the rates under the no group difference condition (Table 7); the former rates were much higher than the latter rates. On the contrary, the power rates under the group SD difference were lower than those under the no group difference condition.

A similar pattern was found for the logistic regression procedure. The Type I rates between the group mean difference condition and the no group difference condition were similar. The power rates between the two conditions were also similar. However, the Type I error rates were higher under the group SD difference condition than under the no group difference condition; the power rates were lower under the former condition (Table 7 and Table 8).

The Type of I error rates of the Lord's chi-square test between the group mean difference condition and the no group difference condition were roughly the same.

However, the rate under the group SD difference condition for the 40 items test was higher than its counterpart under the no group difference condition (Table 7). The power rates of the Lord's chi-square test were similar for the two conditions: group mean difference and no group difference conditions. On the other hand, the power rates under the group SD difference condition were lower than those under the no group difference condition (Table 8).

While the MH method, the logistic regression procedure, and the Lord's chi-square test were all affected by the group SD difference conditions discussed above, the DFIT method was not affected much by any group difference conditions. In particular, in the 20 items test, the Type I error rates were the same for all three conditions (.06) (Table 7). In the 40 items test, the Type I error rate under the group SD difference condition were slightly higher than the rate under the no group difference condition. The power rates under the group SD difference condition dropped slightly compared to those under the no group difference condition. Therefore, the DFIT method showed the most stable performance across the various group difference conditions.

In summary, the DFIT method seemed to be the most effective method for controlling the Type I error rate, especially when the group SD difference existed because it did not inflate the Type I error rate as much as the other methods. Across all conditions of sample size, sample size ratio, test length, and group difference, the DFIT method generally performed better than the other three methods[1]. Especially the DFIT method performed very well under the condition of group SD difference compared to the other

---

[1]For the 20-item test, the Lord's chi-square test exhibited the lowest the Type I error rates.

methods. The finding presented here suggests that the group SD difference condition inflates of the Type I error rate. Figure 6 and Figure 7 show the inflated Type I error rates, under the group SD difference condition, of the four DIF methods in the 20 items test and for the 40 items test, respectively.

The study results also suggest that high power rates were achieved with large DIF magnitude items in general (Table 8). The power rates were lower when groups' SDs were different than when there was no group difference or than when group means were different.

Table 7

Type I error rate for four DIF methods by group difference

Test Length	Sample size	Group Diff.	Type I error rate			
			MH method	Logistic Regression	DFIT	Lord's chi-square
20 items test	500/500	No	.13	.13	.06	.03
	1000/1000		.19	.20		
	700/300		.09	.10		
	1500/500		.11	.12		
	500/500	Mean	.11	.12	.06	.03
	1000/1000		.16	.18		
	700/300		.10	.11		
	1500/500		.09	.10		
	500/500	SD	.23	.71	.06	.33
	1000/1000		.36	.84		
	700/300		.20	.67		
	1500/500		.22	.63		
40 items Test	500/500	No	.12	.13	.13	.13
	1000/1000		.17	.18		
	700/300		.12	.13		
	1500/500		.09	.11		
	500/500	Mean	.11	.13	.13	.14
	1000/1000		.16	.17		
	700/300		.11	.12		
	1500/500		.10	.11		
	500/500	SD	.16	.64	.15	.20
	1000/1000		.25	.78		
	700/300		.16	.64		
	1500/500		.13	.51		

Table 8

The power rate for four DIF methods by group difference

Test Length	Sample size	Group Diff.	Power Rate (With DIF magnitude)			
			MH method	Logistic Regression	DFIT	Lord's chi-square
20 items test	500/500	No	1.00 (1)	1.00 (1)	.95 (1)	1.00 (1)
	1000/1000		1.00 (1)	1.00 (1)		
	700/300		.98 (1)	1.00 (1)		
	1500/500		1.00 (1)	1.00 (1)		
	500/500	Mean	1.00 (1)	1.00 (1)	.96 (1)	.99 (1)
	1000/1000		1.00 (1)	1.00 (1)		
	700/300		1.00 (1)	1.00 (1)		
	1500/500		1.00 (1)	1.00 (1)		
	500/500	SD	.87 (1)	.62 (1)	.84 (1)	.48 (1)
	1000/1000		.97 (1)	.66 (1)		
	700/300		.89 (1)	.60 (1)		
	1500/500		.78 (1)	.56 (1)		
40 items Test	500/500	No	1.00 (1)	1.00 (1)	.99 (1)	.99 (1)
			.28 (.5)	.30 (.5)	.49 (.5)	.55 (.5)
	1000/1000		1.00 (1)	1.00 (1)		
			.48 (.5)	.49 (.5)		
	700/300		1.00 (1)	1.00 (1)		
			24 (.5)	.28 (.5)		
	1500/500		.95 (1)	.97 (1)		
			.23 (.5)	.21 (.5)		
	500/500	Mean	1.00 (1)	1.00 (1)	.97 (1)	.99 (1)
			.33 (.5)	.33 (.5)	.35 (.5)	.50 (.5)
	1000/1000		1.00 (1)	1.00 (1)		
			.49 (.5)	.48 (.5)		
	700/300		1.00 (1)	1.00 (1)		
			.29 (.5)	.31 (.5)		
	1500/500		.97 (1)	.98 (1)		
			.23 (.5)	.21 (.5)		
	500/500	SD	.89 (1)	.60 (1)	.81 (1)	.58 (1)
			.09 (.5)	.66 (.5)	.53 (.5)	.52 (.5)
	1000/1000		.98 (1)	.62 (1)		
			.16 (.5)	.89 (.5)		
	700/300		.79 (1)	.51 (1)		
			.08 (.5)	.74 (.5)		
	1500/500		.67 (1)	.46 (1)		
			.11 (.5)	.59 (.5)		

*(1: large) and (.5: medium) indicate DIF magnitude (the *b* item difficulty parameter difference between reference and focal group)

Figure 6

Inflated Type I error rate with group SD difference for four DIF methods in a 20 items *test*

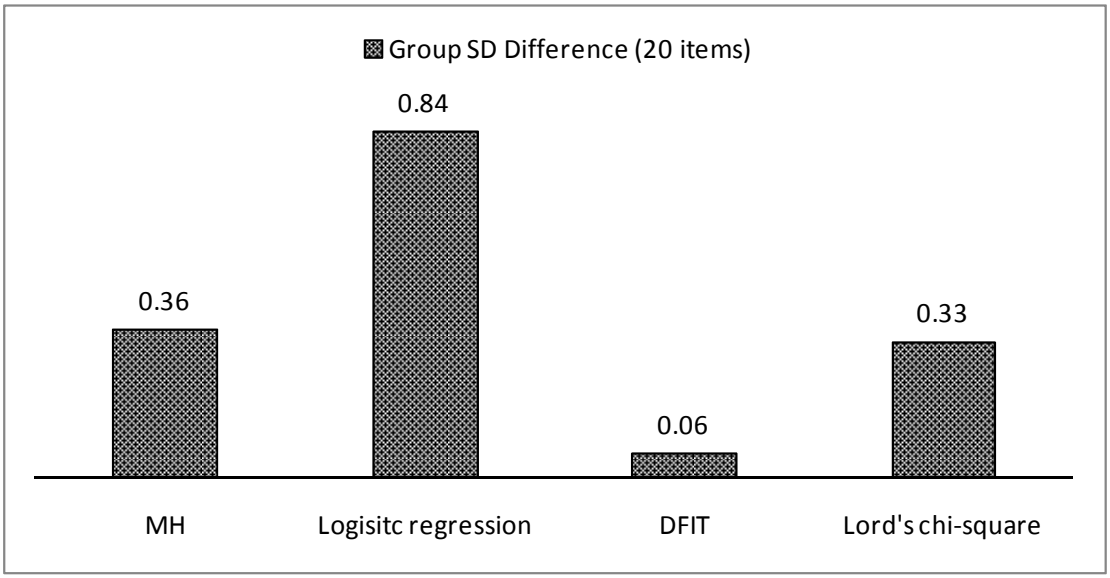
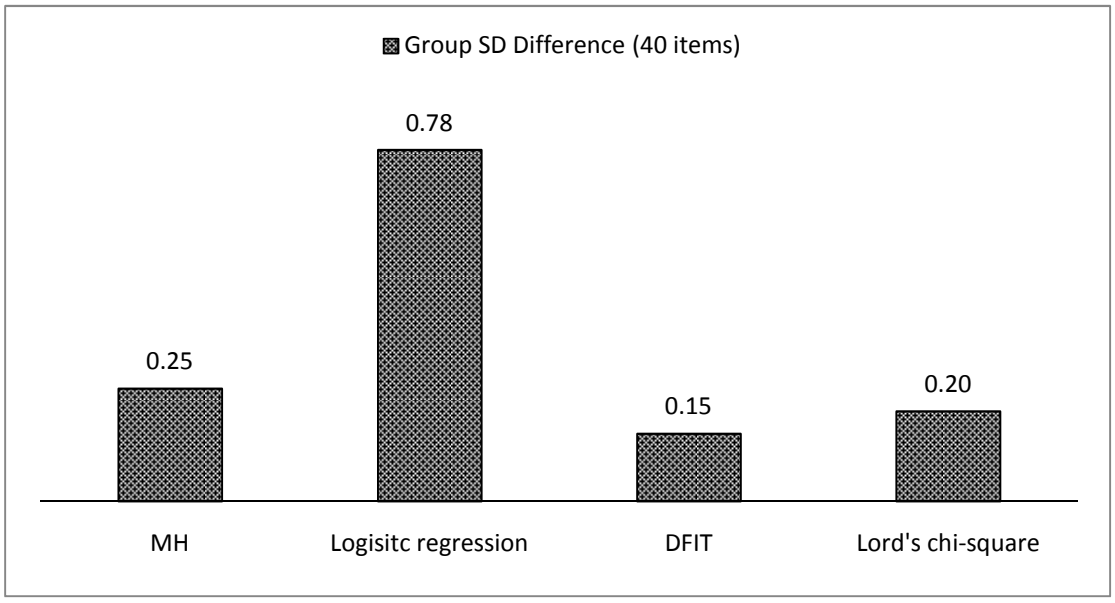


Figure 7

Inflated Type I error rate with a group SD difference for four DIF methods in a 40 items test



Phase II

The second set of research questions—2a, 2b, and 2c—asks about the effectiveness of the three adjustment procedures in controlling the Type I error rate of the four DIF methods. This section answers these research questions.

The findings in this section indicate that all three adjustment procedures reduced the Type I error rates effectively. One negative effect of using such procedures, however, was the decreased power rates. Of the three adjustment procedures, The BH method seemed to be the most balanced method—reducing Type I error rates while not losing the power as much as the other methods did.

Effect of the three adjustment procedures on the MH method

Table 9 shows Type I error rates and power rates of the MH method in the 20 items test, both unadjusted and adjusted (i.e., the Bonferroni correction, the Holm's procedure, and BH method), broken down by the group difference conditions (i.e., no group difference, group mean difference, and group SD difference). Most unadjusted Type I error rates were above .10 (or 10%) with the maximum rate excess of .30. All three adjustment procedures lowered the Type I error rates effectively—below .10, and in most case below .05. Such effectiveness was somewhat withered under the group SD difference condition, however. Even though all three procedures reduced the power rates, the BH method seemed to lose power the least. Note that both unadjusted and adjusted power rates were high due to the large DIF magnitude used in the 20 items test.

Table 9

Type I error rate/ power for the MH method with group difference (20 items test)

MH Method									
Condition		Type I Error Rate				Power Rate (With DIF magnitude)			
Group Diff.	Sample Size	Unadj.	Bonf.	Holm's	BH	Unadj.	Bonf.	Holm's	BH
No	500/500	.13	.01	.01	.04	1.00 (1)	.99 (1)	.99 (1)	1.00 (1)
	1000/1000	.19	.02	.02	.08	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)
	700/300	.09	.01	.01	.02	.98 (1)	.91 (1)	.92 (1)	.94 (1)
	1500/500	.11	.01	.01	.03	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)
Mean	500/500	.11	.01	.01	.03	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)
	1000/1000	.16	.02	.02	.06	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)
	700/300	.10	.01	.01	.02	1.00 (1)	.96 (1)	.96 (1)	.97 (1)
	1500/500	.09	.01	.01	.02	1.00 (1)	.95 (1)	.95 (1)	.96 (1)
SD	500/500	.23	.07	.07	.11	.87 (1)	.68 (1)	.69 (1)	.76 (1)
	1000/1000	.36	.13	.13	.25	.97 (1)	.86 (1)	.88 (1)	.96 (1)
	700/300	.20	.05	.05	.09	.89 (1)	.66 (1)	.66 (1)	.72 (1)
	1500/500	.22	.06	.06	.11	.78 (1)	.54 (1)	.54 (1)	.62 (1)

*(1) indicates large DIF magnitude (the difference of *b* item difficulty parameter between the reference and focal group)

Table 10 shows the Type I error rates and power rates of the MH method in the 40 items test, both unadjusted and adjusted, broken down by the group difference conditions. The patterns found were similar to those found in the 20 items test: all three procedures reduced the Type I error rate and the power rates. All three procedures suffered the loss of the power, and such a loss was more severe when the DIF magnitude was medium (i.e., 0.5).

Table 10

Type I error rates / power rates for the MH method with group difference (40 items test)

Condition		MH Method							
		Type I Error Rate				Power Rate(With DIF magnitude)			
Group Diff.	Sample Size	Unadj.	Bonf.	Holm's	BH	Unadj.	Bonf.	Holm's	BH
No	500/500	.12	.02	.02	.04	1.00 (1) .28 (.5)	.96 (1) .05 (.5)	.96 (1) .05 (.5)	.98 (1) .10 (.5)
	1000/1000	.17	.04	.04	.08	1.00 (1) .48 (.5)	1.00 (1) .12 (.5)	1.00 (1) .12 (.5)	1.00 (1) .23 (.5)
	700/300	.12	.02	.02	.04	1.00 (1) .24 (.5)	.94 (1) .03 (.5)	.94 (1) .03 (.5)	.95 (1) .06 (.5)
	1500/500	.09	.01	.01	.02	.95 (1) .23 (.5)	.65 (1) .02 (.5)	.66 (1) .02 (.5)	.74 (1) .04 (.5)
Mean	500/500	.11	.01	.01	.30	1.00 (1) .33 (.5)	.94 (1) .04 (.5)	.94 (1) .04 (.5)	.98 (1) .10 (.5)
	1000/1000	.16	.04	.04	.07	1.00 (1) .49 (.5)	1.00 (1) .13 (.5)	1.00 (1) .15 (.5)	1.00 (1) .24 (.5)
	700/300	.11	.02	.02	.04	1.00 (1) .29 (.5)	.91 (1) .03 (.5)	.91 (1) .03 (.5)	.96 (1) .09 (.5)
	1500/500	.10	.01	.01	.02	.97 (1) .23 (.5)	.61 (1) .02 (.5)	.61 (1) .02 (.5)	.69 (1) .03 (.5)
SD	500/500	.16	.03	.03	.05	.89 (1) .09 (.5)	.43 (1) .00 (.5)	.45 (1) .00 (.5)	.63 (1) .00 (.5)
	1000/1000	.25	.07	.07	.12	.98 (1) .16 (.5)	.78 (1) .01 (.5)	.79 (1) .01 (.5)	.92 (1) .03 (.5)
	700/300	.16	.04	.04	.06	.79 (1) .08 (.5)	.43 (1) .05 (.5)	.44 (1) .05 (.5)	.58 (1) .15 (.5)
	1500/500	.13	.02	.02	.03	.67 (1) .11 (.5)	.21 (1) .00 (.5)	.22 (1) .00 (.5)	.26 (1) .01 (.5)

*(1: large) and (.5: medium) indicate DIF magnitude (the *b* item difficulty parameter difference between reference and focal group)

Effect of the three adjustment procedures on the logistic regression

Table 11and Table 12 display the Type I error rates and power rates before and after the adjustments for the logistic regression procedure in the 20 items and 40 items tests, respectively. The general patterns found from these tables were similar to those found in the prior section: all three procedures reduced both the Type I error rates and the power rates. The main difference between the MH method and the logistic regression was in the group SD difference condition. The logistic regression produced even higher inflated Type I error rates and more deflated power rates under the group SD difference condition. All three procedures reduced the Type I error rates under the group SD difference condition, but adjusted Type I error rates were still too high.

Another interesting pattern was found in the power rates under the group SD difference condition. For unadjusted rates, the power rates for the medium DIF magnitude items were generally greater than their counterpart for the large DIF magnitude items (Table 12). For example, the unadjusted power rate for the medium DIF magnitude items with 1000/1000 sample size was .89, and its counterpart in the large DIF magnitude items was .62. However, such a pattern was often reversed when the Bonferroni correction or the Holm's procedure was applied. The BH method preserved the original pattern.

Table 11

Type I error rate/ power rate for logistic regression procedure with group difference (20 items test)

Logistic Regression									
Condition		Type I Error Rate				Power Rate (With DIF magnitude)			
Group Diff.	Sample Size	Unadj.	Bonf.	Holm's	BH	Unadj.	Bonf.	Holm's	BH
No	500/500	.13	.01	.02	.04	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)
	1000/1000	.20	.03	.03	.08	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)
	700/300	.10	.01	.01	.02	1.00 (1)	.95 (1)	.95 (1)	.95 (1)
	1500/500	.12	.01	.01	.03	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)
Mean	500/500	.12	.01	.01	.03	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)
	1000/1000	.18	.03	.03	.06	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)
	700/300	.11	.01	.01	.02	1.00 (1)	.97 (1)	.97 (1)	.99 (1)
	1500/500	.10	.01	.01	.03	1.00 (1)	.98 (1)	.98 (1)	.99 (1)
SD	500/500	.71	.50	.54	.68	.62 (1)	.50 (1)	.52 (1)	.59 (1)
	1000/1000	.84	.70	.75	.84	.66 (1)	.54 (1)	.59 (1)	.65 (1)
	700/300	.67	.42	.46	.63	.60 (1)	.48 (1)	.51 (1)	.58 (1)
	1500/500	.63	.40	.43	.59	.56 (1)	.47 (1)	.49 (1)	.55 (1)

*(1: large) and (.5: Medium) indicate DIF magnitude (the *b* item difficulty parameter difference between reference and focal group)

Table 12

Type I error rate / power for logistic regression procedure with group difference (40 items test)

Condition		Logistic Regression Procedure							
		Type I Error Rate				Power Rate (With DIF magnitude)			
Group Diff.	Sample Size	Unadj.	Bonf.	Holm's	BH	Unadj.	Bonf.	Holm's	BH
No	500/500	.13	.04	.04	.05	1.00 (1) .30 (.5)	.98 (1) .06 (.5)	.98 (1) .07 (.5)	.99 (1) .11(.5)
	1000/1000	.18	.06	.06	.09	1.00 (1) .49 (.5)	1.00 (1) .19 (.5)	1.00 (1) .19 (.5)	1.00 (1) .31 (.5)
	700/300	.13	.03	.04	.06	1.00 (1) .28 (.5)	.96 (1) .04 (.5)	.96 (1) .04 (.5)	.99 (1) .08 (.5)
	1500/500	.11	.02	.02	.03	.97 (1) .21 (.5)	.75 (1) .01 (.5)	.75 (1) .01 (.5)	.83 (1) .04 (.5)
Mean	500/500	.13	.03	.03	.05	1.00 (1) .33 (.5)	.98 (1) .04 (.5)	.98 (1) .04 (.5)	1.00 (1) .12 (.5)
	1000/1000	.17	.05	.05	.08	1.00 (1) .48 (.5)	1.00 (1) .17 (.5)	1.00 (1) .18 (.5)	1.00 (1) .28 (.5)
	700/300	.12	.03	.03	.05	1.00 (1) .31 (.5)	.97 (1) .03 (.5)	.97 (1) .03 (.5)	.99 (1) .08 (.5)
	1500/500	.11	.01	.01	.02	.98 (1) .21 (.5)	.71 (1) .02 (.5)	.71 (1) .02 (.5)	.76 (1) .05 (.5)
SD	500/500	.64	.37	.40	.60	.60 (1) .66 (.5)	.36 (1) .27 (.5)	.37 (1) .31 (.5)	.56 (1) .60 (.5)
	1000/1000	.78	.58	.63	.76	.62 (1) .89 (.5)	.48 (1) .62 (.5)	.52 (1) .72 (.5)	.61 (1) .87 (.5)
	700/300	.64	.36	.39	.59	.54 (1) .74 (.5)	.36 (1) .25 (.5)	.37 (1) .29 (.5)	.52 (1) .67 (.5)
	1500/500	.51	.25	.26	.43	.46 (1) .59 (.5)	.17 (1) .15 (.5)	.20 (1) .19 (.5)	.41 (1) .45 (.5)

*(1: large) and (.5: Medium) indicate DIF magnitude (the *b* item difficulty parameter difference between reference and focal group)

The Effect of the three adjustment procedures on the DFIT method

Table 13 displays the Type I error rates and power rates before and after the adjustments for the DFIT method in the 20 items and 40 items tests. According to the table, all three adjustment procedures worked well with the DFIT method; that is, they reduced the Type I error rates. One important thing to note is that the DFIT method had controlled the Type I error rate very well even before the adjustment procedure was applied in the 20 items test. It means that the DFIT method did not need the adjustment when the test length was short (e.g., 20 items test). Not having to adjust is beneficial because adjustment will lower the power rates. However, in the 40 items test, the unadjusted Type I error rates were high and the adjustment procedures lowered the rate substantially. As with the adjustment procedures discussed earlier, the adjusted power rates were lower than unadjusted counterparts. The unadjusted power rates under the group SD difference condition were generally lower—but not by much—than those under the other group difference conditions. The reduction in the power rate after adjustment was the biggest under the group SD difference condition.

Table 13

The effect of three adjustment procedures for the DFIT method with group difference

Condition		DFIT							
		Type I Error Rate				Power Rate (With DIF magnitude)			
Group Diff.	Test length	Unadj.	Bonf.	Holm's	BH	Unadj.	Bonf.	Holm's	BH
No	20	.06	.01	.00	.01	.95	.65	.48	.69
Mean		.06	.01	.00	.02	.96	.69	.47	.71
SD		.06	.01	.01	.02	.84	.42	.41	.47
No	40	.13	.04	.04	.06	.99 (1) .49 (.5)	.73 (1) .18 (.5)	.73 (1) .18 (.5)	.79 (1) .34 (.5)
Mean		.13	.02	.00	.04	.97 (1) .35 (.5)	.66 (1) .12 (.5)	.49 (1) .05 (.5)	.72 (1) .21 (.5)
SD		.15	.04	.02	.06	.81 (1) .53 (.5)	.28 (1) .11 (.5)	.17 (1) .05 (.5)	.44 (1) .19 (.5)

*(1: large) and (.5: medium) indicate DIF magnitude (the *b* item difficulty parameter difference between reference and focal group)

Effect of the three adjustment procedures on the Lord's chi-square test

The Lord's chi-square test showed results containing features of both non-parametric and parametric methods. As with the DFIT method, the unadjusted Type I error rates of the Lord's chi-square test in a short length test (e.g., 20 items test) were well-controlled for either no group difference or group mean difference condition. However, the Lord's chi-square test turned out to be influenced by the group SD difference condition like the non-parametric methods (the MH method and the logistic regression procedure), and all three adjusted Type I error rates still showed slight inflation (Table 14).

Table 14

Effect of three adjustment procedures for Lord's chi-square test with group difference

Lord's Chi-square Test									
Condition		Type I Error Rate				Power Rate (With DIF magnitude)			
Group Diff.	Test length	Unadj.	Bonf.	Holm's	BH	Unadj.	Bonf.	Holm's	BH
No	20	.03	.00	.00	.01	1.00	.98	.98	.99
Mean	20	.03	.00	.00	.01	.99	.94	.94	.97
SD	20	.33	.18	.19	.26	.48	.29	.30	.39
No	40	.13	.05	.05	.07	.99 (1) .55 (.5)	.93 (1) .17 (.5)	.93 (1) .18 (.5)	.97 (1) .32 (.5)
Mean	40	.14	.05	.05	.08	.99 (1) .50 (.5)	.88 (1) .11 (.5)	.89 (1) .11 (.5)	.96 (1) .25 (.5)
SD	40	.20	.10	.10	.13	.58 (1) .52 (.5)	.51 (1) .47 (.5)	.51 (1) .47 (.5)	.54 (1) .50 (.5)

*(1: large) and (.5: medium) indicate DIF magnitude (the *b* item difficulty parameter difference between reference and focal group)

CHAPTER 5

DISCUSSION

Conclusion and Significance

The simulation in this study had two phases, each of which answered the research questions constructed in Chapter 1. The first was to investigate the influence of various factors—sample size, test length, and group difference—on the Type I error rates of the four DIF methods, and the second was to apply adjustment procedures to lower the Type I error rates to the case of multiple significance tests.

The findings of the first phase revealed that all testing conditions considered in this study influenced the Type I error rates of both non-parametric methods and parametric methods. In terms of test length, both non-parametric and parametric methods performed similarly in the 40 items test. In the 20 items test, however, the Type I error rates of the non-parametric methods were higher than those of the parametric methods. Also, the effect of the test length on the Type I error rates of the parametric methods were significant: a longer test (e.g., 40 items test) inflated the Type I error rate more than a shorter test (e.g., 20 items test). The effect of the test length was not significant for the non-parametric methods.

The results also showed that large sample size and equal ratio of sample size tended to inflate the Type I error rates of all four DIF methods. When the condition of the group difference was concerned, the presence of the group mean difference did not

influence the Type I error rates much. However, the presence of the group SD difference significantly inflated the Type I error rates of all but the DFIT method. This is one of valuable findings of this study.

In sum, the results of the first phase simulation suggested that the DFIT method was the most effective method to control the Type I error rates under the testing conditions considered in this study.

The results also showed the trend of the power rate. As explained, in the simulation study, the two different levels of DIF magnitude were set up. With the large DIF magnitude, the power rates were consistently high—close to 1. However, the power rates were comparably low with medium the DIF magnitude.

The findings of the second phase simulation answered the research questions asking about the effectiveness of adjustments in controlling the Type I error rate. All three procedures—the Bonferroni correction, the Holm's procedure and the BH method—were effective in controlling the Type I error rates of all four DIF methods. For the non-parametric methods, the adjustment procedures reduced the Type I error rates, except for the condition of the group SD difference, even though the power rates were also reduced. Of three adjustment procedures, the BH method seemed to be the most balanced method in lowering the Type I error rate and at the same time not losing too much power, compared to the Bonferroni correction and the Holms procedure.

One interesting finding was that when the test length was short (e.g., 20 items test) the Type I error rate of the DFIT method and of the Lord's chi-square test were well-controlled even before adjustment. Therefore, for the parametric tests investigated here, adjustment may not be necessary for a shorter test, but the benefit of adjustment may

increase as the test length becomes longer. On the other hand, for the non-parametric tests investigated here, adjustment may be beneficial at any test with any test length. In sum, adjustment procedures were effective in controlling the Type I error rate in DIF analysis. This finding is invaluable in DIF studies because the issues of multiple significance testing, which have been studied quite often in applied statistics, have been rarely studied in DIF research. This study serves as one of the front runners in and at the same time as a fodder for future research in adjustment of multiple significance tests in DIF studies.

Limitations and Future Research

The study found a relationship between the inflation of the Type I error rate and several testing conditions. In particular, two major findings deserve further investigation. First, the effect of group SD difference on the inflation of the Type I error rate was distinct. This dissertation found this phenomenon through the simulation study. The next step would be to verify and determine why and how much the group SD difference causes serious inflation of the Type I error rate.

Second, the findings showed that the Type I error rate of the DFIT method was not affected by the group SD difference condition. Therefore, further research should be conducted to explore the factors that enable the DFIT method to control the Type I error rate consistently in any particular condition, such that the DFIT method is based on an empirical distribution.

Even though the group SD difference always exists in the practical testing fields, research on the former difference has been neglected so far compared to the group mean

difference. Therefore, further research on the influence of the group SD difference on the Type I error rate is warranted.

This dissertation examined the effectiveness of Type I error rates and the adjustment procedure for multiple testing for only dichotomous items. Further research should assess the effectiveness of Type I error rates and adjustment procedure for multiple testing for polytomous items as well.

References

Agresti, A. (1996). *An Introduction to categorical data analysis*. Hohn Wiley & Sons, Inc.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289-300.

Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement*, *37*(4), 307-327.

Bradley, J. V. (1978). Robustness? *The British Journal of Mathematical & Statistical Psycoholgy*, *31*, 144-152.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics*, *18*(2), 131-154.

Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement*, *68*(6), 940-958.

Finch, W. H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT ikelihood Ratio. *Applied Psychological*

Measurement, *29*(4), 278-295.

Gierl, M. L., Gotzmann, A., & Boughton, K. A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education*, *17*(3), 241-264.

Gotzmann, A., & Boughton, K. A. (2004). *A comparison of type I error and power for the Mantel Haenszel and SIBTEST procedures when group differences are large and unbalanced*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Measurement Methods for the Social Sciences Series (Sage Publications.). Newbury Park, London. Retrieved from http://scholar.google.com/scholar?start=10&q=Lord%27s+Chi+square&hl=en

Holland, B. S., & Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics*, *43*, 417-423.

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65-70.

Jodoin, M. G., & Gierl, M. L. (2001). Evaluating Type I error and Power rates using an effect size measure with the Logistic Regression procedure for DIF detection. *Applied Measurement in Education*, *14*, 329-349.

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

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). *A review of recent developments in differential item functioning* (Research Report No. ETS RR-08-43). Princeton, NJ: ETS.

Minium, E. W., Clarke, R. C., & Coladarci, T. (1998). *Elements of Statistical Reasoning* (2nd ed.). John Wiley & Sons, Inc.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*(4), 315-328.

Oshima, T., Raju, N., & Nanda, A. O. (2006). A new Method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement*, *43*(1), 1-17.

Penny, J., & Johnson, R. L. (1999). How group differences in matching criterion distribution and IRT item difficulty can influence the magnitude of the Mantel Haenszel chi-square DIF index. *Journal of Experimental Education*, *67*(4), 343-

367.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments? *British Medical Journal*, *316*(7139), 1236-1238.

Raju, N., Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353-368.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116.

Roussos, L. A., & Stout, W. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel- Haenszel Type I Error performance. *Journal of Educational Measurement*, *33*(2), 215-230.

Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. *The American Statistician*, *40*(4), 313-315.

Sheppard, R., Han, K., Colarelli, S., Dai, G., & King, D. W. (2006). Differential item functioning by sex and race in the hogan personality inventory. *Assessment*, *13*(4), 442-453.

Soulakova, J. N. (2009). On identifying effective and superior drug combinations via Holm's procedure based on the Min tests. *Journal of Biopharmaceutical Statistics*,

19(2), 280-291.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology*, *81*(2), 332-342.

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

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, *27*(1), 77-83.

Williams, V., Jones, L., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state to state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, *24*, 42-69.

Zumbo, B. D. (1999). *A handbook on theory and methods of differential item functioning (DIF)* (No. K1A 0K2) (pp. 1-57). Ottawa, Canada: Human Resources Research and Evaluation, National Defense.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*(3), 233-251.

APPENDIXES

APPENDIX A

Classification of statistical methods based on Interpretable measure of amount of DIF (Mapuranga, Dorans, & Middleton, 2008)

Statistical Methods	Measures of amount of DIF
Logistic Regression	P-DIF, R^2-like indices and log odds ratios
Hierarchical logistic regression	Effect size estimate based on log odds ratio or R^2-like index based on change in variance components
Logistic mixed model	Log odds ratio estimate conditional on latent ability
Mixture model	Likelihood ratio
HGLM	Log odds ratio
DFIT	Compensatory (CDIF) & non compensatory (NCDIF) DIF statistics
TestGraf	Root mean square average difference
Scrams McLeod	MH statistic
Mantel Haenszel	MH log odds ratio estimate
Cochran Mantel Haenszel	A set of odds ratios
Liu-Agresti estimator	Liu-Agresti cumulative common odds ratio
Cox's β	Log odds ratios

APPENDIX B

Classification of statistical methods based on index or test of significance (Mapuranga et al., 2008)

Statistical Methods	Index or test of significance based
DFIT	Chi square & t tests of significance
MIMIC model	Significance test of model coefficients
Lagrangian multiplier tests	Lagrangian multiplier test of significance
RCML	Hotelling's T
Logistic Regression	Chi square test of significance
Hierarchical logistic regression	Significance test of model coefficients
Logistic mixed model	Wald test of significance
Mixture model	Chi square test of significance
HGLM	Chi-square significance test of model coefficients
Mantel Haenszel	MH chi square test of significance
Cochran Mantel Haenszel	Chi-square significance test
Liu-Agresti estimator	Liu-Agresti cumulative common odds ratio
Cox's β	Cox's β test statistic
SIBTEST	SIB test statistic

APPENDIX C

IPR Procedure (Oshima et al., 2006)

The IPR procedure proposed new cutoff values for each item by $(1-\alpha)$ percentile rank score from a frequency distribution of NCDIF values under the no DIF conditions in the DFIT framework. The fixed cutoff value of NCDIF index for an item, *i,* (Raju, van der Linden, & Fleer, 1995) is defined as .006 in the DFIT framework for dichotomous item analysis (Fleer, 1993). In order to improve the procedure for assessing the statistical significance of NCDIF index, the IPR method developed cutoff values ranging from .003 to .15 with different conditions of sample size and IRT model, which conditions are "with a higher cutoff value for a smaller sample size and a higher value for an IRT model with more parameters" (p. 2). The advantage of the IPR procedure is to "omit the whole process of the steps from item parameter calibration, linking, to analysis" (2006, p. 7). And, the IPR procedure can generate a cutoff value for each item, unlike DFIT that produces one cutoff value for all items. The IPR procedure has a computer program developed, which is called "DIFCUT" (Nanda, Oshima, & Gagne, 2006), so that practitioners can do DIF analysis easily. The algorithm of the IPR procedure is below (p. 4).

1. Let the item parameter estimates from the focal group be denoted by a column vector, M_i, for item *i*. In the case of the 3PL model, M_i will consist of 3 elements (b_i, a_i, and c_i item parameters) as shown below:

$$M_i = \begin{bmatrix} b_i \\ a_i \\ c_i \end{bmatrix}.$$

In the case of the 1PL or the Rasch model, M_i will be a scalar with an estimate of the *b* parameter. Associated with each item is a matrix, V_i, consisting of the sampling variances and covariances of the item parameter estimates:

$$V_i = \begin{bmatrix} \sigma^2_{b_i} & \sigma_{b_i a_i} & \sigma_{b_i c_i} \\ \sigma_{a_i b_i} & \sigma^2_{a_i} & \sigma_{a_i c_i} \\ \sigma_{c_i b_i} & \sigma_{c_i a_i} & \sigma^2_{c_i} \end{bmatrix}.$$

The information in V_i is also typically provided by the commercially available IRT calibration programs. Let R_i represent the correlation matrix for the item parameters of item *i*. These item parameter inter-correlations can be derived from V_i:

$$R_i = \begin{bmatrix} 1 & \rho_{b_i a_i} & \rho_{b_i c_i} \\ \rho_{a_i b_i} & 1 & \rho_{a_i c_i} \\ \rho_{c_i b_i} & \rho_{c_i a_i} & 1 \end{bmatrix}.$$

Assuming that R_i is positive definite, it can be expressed as the product of a triangular matrix (T_i) and its transpose (T_i') (Graybill, 1969); that is:

$$R_i = T_i' T_i .$$

In the present context, T_i can be expressed as:

$$T_i = \begin{bmatrix} 1 & \rho_{b_i a_i} & \rho_{b_i c_i} \\ 0 & \sqrt{1-\rho_{b_i a_i}^2} & \dfrac{\rho_{a_i c_i} - \rho_{b_i a_i}\rho_{b_i c_i}}{\sqrt{1-\rho_{b_i a_i}^2}} \\ 0 & 0 & \sqrt{1-\left[\rho_{b_i c_i}^2 + \dfrac{(\rho_{a_i c_i} - \rho_{b_i a_i}\rho_{b_i c_i})^2}{(1-\rho_{b_i a_i}^2)}\right]} \end{bmatrix}.$$

For the 2PL model, the above matrix reduces to:

$$T_i = \begin{bmatrix} 1 & \rho_{b_i a_i} \\ 0 & \sqrt{1-\rho_{b_i a_i}^2} \end{bmatrix}.$$

For the Rasch model, T_i becomes a scalar with a unit as its value.

2. Let k represent the IRT model under consideration. For the Rasch model, $k = 1$, for 2PL, $k = 2$, and for 3PL, $k = 3$. Now, let X_{1i} represent a column vector of k elements, with each element drawn at random from one of k independent, standardized (mean of 0 and standard deviation of 1), and normally distributed populations. Let X_{2i} represent a second vector of k elements similarly drawn.
3. Using the T_i matrix in Equation 9, transform the two X vectors into two Z (column) vectors as follows:

$$Z_{1i} = T'_i X_{1i},$$

$$Z_{2i} = T'_i X_{2i}.$$

Each Z vector now represents a random element from a k-dimensional standardized multivariate normal distribution with a correlation structure for the k dimensions conforming to the correlation structure in the R_i matrix.

4. By definition, each element in the Z vectors is standardized in that its expectation and variance are 0 and 1, respectively. Each Z vector is now transformed to a Y vector so that the elements in the new vector will have the appropriate mean and variance as shown in the M_i and V_i matrices above. To achieve this transformation, let D_i represent a diagonal matrix consisting of the diagonal elements (variances) in V_i. Now, let

$$Y_{1i} = D_i^{1/2} Z_{1i} + M_i ,$$

$$Y_{2i} = D_i^{1/2} Z_{2i} + M_i .$$

5. Vectors Y_{1i} and Y_{2i} represent two estimates of item parameters from two populations with identical item parameters; these vectors may be thought as representing item parameter estimates for the focal and reference groups when the true DIF is zero. That is, any difference in these two sets of estimates is simply due to sampling error. Therefore, an NCDIF index for item i can be obtained with the help of the two Y vectors and the estimates of thetas for the focal group, using the computations spelled out in Raju et al. (1995).
6. Steps 1-5 can be replicated as many times as one wishes (for example, 100, 1000, …, or 10,000 times).

7. NCDIF values from all replications obtained in Step 6 will be rank ordered and the 90th, 95th, 99th, and 99.9th percentile rank scores are recorded to establish the cutoff values for alpha levels at .10, .05, .01, and .001, respectively.
8. Once the alpha level is chosen, the cutoff associated with it will be used as the cutoff for assessing statistical significance of the initial NCDIF value obtained for item *i*.
9. Steps 1-8 are repeated for all items in the test, thus potentially resulting in different cutoffs for different items.

   A SAS-IML program "DIFCUT" is used to process the algorithm above. This IPR procedure can be used in all 1PL, 2PL, and 3PL IRT model with dichotomous items basis.

APPENDIX D

DETECTION OF FALSE POSITIVES IN MANTEL HAENSZEL METHOD BEFORE AND AFTER THREE ADJUSTMENTS

Condition 1 (20 items test length, 500/500, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	11	0	0	1
100	2	15	3	4	6
100	3	100	98	98	99
100	4	15	1	1	6
100	5	5	0	0	1
100	6	5	0	0	2
100	7	22	2	2	5
100	8	100	100	100	100
100	9	17	5	5	6
100	10	13	1	1	4
100	11	16	0	0	6
100	12	12	1	3	7
100	13	8	1	1	1
100	14	10	0	0	3
100	15	7	0	1	1
100	16	16	3	3	4
100	17	21	1	1	10
100	18	18	4	4	6
100	19	13	0	0	3
100	20	14	2	2	3

Condition 2 (20 items test length, 1000/1000, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	17	2	2	6
100	2	12	2	2	5
100	3	100	100	100	100
100	4	15	2	3	4
100	5	11	2	2	5
100	6	13	1	2	8
100	7	18	4	4	9
100	8	100	100	100	100
100	9	22	2	3	7
100	10	16	1	1	5
100	11	28	3	3	13
100	12	24	4	4	12
100	13	15	2	2	6
100	14	11	2	2	3
100	15	10	0	0	3
100	16	29	2	2	13
100	17	34	6	6	14
100	18	28	4	4	14
100	19	12	2	2	7
100	20	21	3	3	11

Condition 3 (20 items test length, 700/300, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	17	0	0	0
100	2	10	1	2	4
100	3	100	99	99	100
100	4	9	0	0	1
100	5	9	0	0	1
100	6	7	1	1	2
100	7	9	1	1	3
100	8	100	100	100	100
100	9	11	3	3	6
100	10	17	2	2	2
100	11	10	1	1	4
100	12	9	0	1	2
100	13	10	1	1	3
100	14	7	0	0	2
100	15	4	0	0	0
100	16	12	2	2	3
100	17	12	1	1	4
100	18	18	4	5	9
100	19	10	0	0	1
100	20	21	0	0	5

Condition 4 (20 items test length, 1500/500, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	6	1	1	3
100	2	9	0	0	1
100	3	96	83	84	89
100	4	8	0	0	0
100	5	11	0	0	1
100	6	6	1	1	1
100	7	6	1	1	1
100	8	100	99	99	99
100	9	13	0	0	1
100	10	9	0	0	0
100	11	15	2	2	5
100	12	10	0	0	1
100	13	6	0	0	0
100	14	9	3	3	3
100	15	8	0	0	1
100	16	9	1	1	2
100	17	10	0	0	1
100	18	5	1	1	2
100	19	10	1	1	2
100	20	14	0	0	3

Condition 5 (20 items test length, 500/500, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	6	1	1	1
100	2	9	1	1	4
100	3	100	99	99	100
100	4	12	0	0	2
100	5	7	1	1	1
100	6	8	1	1	1
100	7	12	2	0.02	3
100	8	100	100	100	100
100	9	14	2	2	4
100	10	13	0	0	1
100	11	17	2	3	8
100	12	14	2	2	3
100	13	12	0	0	1
100	14	8	0	0	1
100	15	7	0	0	0
100	16	12	1	1	4
100	17	8	1	1	3
100	18	16	1	1	6
100	19	12	0	0	3
100	20	13	2	2	5

Condition 6 (20 items test length, 1000/1000, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	15	2	2	5
100	2	13	5	5	8
100	3	100	100	100	100
100	4	16	3	3	5
100	5	15	1	1	5
100	6	13	1	1	3
100	7	14	3	3	9
100	8	100	100	100	100
100	9	20	4	4	7
100	10	16	1	1	3
100	11	18	2	2	7
100	12	23	3	3	6
100	13	24	2	2	10
100	14	8	1	1	2
100	15	7	1	1	3
100	16	18	1	1	4
100	17	19	4	4	10
100	18	22	3	4	9
100	19	14	1	1	2
100	20	21	1	2	6

Condition 7 (20 items test length, 700/300, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	7	1	1	1
100	2	9	1	1	2
100	3	99	90	90	91
100	4	14	1	1	1
100	5	8	1	1	2
100	6	5	0	0	0
100	7	12	1	1	4
100	8	100	100	100	100
100	9	11	0	0	2
100	10	14	0	0	3
100	11	4	0	0	1
100	12	5	0	0	1
100	13	4	0	0	0
100	14	10	2	2	5
100	15	5	1	1	2
100	16	9	0	0	1
100	17	6	1	1	1
100	18	13	2	2	5
100	19	11	0	0	4
100	20	13	0	0	5

Condition 8 (20 items test length, 1500/500, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	7	0	0	2
100	2	5	0	0	0
100	3	100	92	92	94
100	4	7	1	1	2
100	5	8	2	2	3
100	6	10	0	0	2
100	7	9	0	0	3
100	8	100	100	100	100
100	9	9	0	0	2
100	10	10	2	2	2
100	11	17	4	4	6
100	12	13	0	0	3
100	13	7	1	1	1
100	14	7	0	0	0
100	15	9	1	1	1
100	16	9	1	1	1
100	17	8	1	1	1
100	18	9	0	0	1
100	19	13	2	2	6
100	20	18	0	0	1

Condition 9(20 items test length, 500/500, and group SD difference*)*

Replication	item	Before	Bonferroni	Holms	BH
100	1	9	0	0	1
100	2	17	2	2	6
100	3	73	36	38	52
100	4	10	0	0	1
100	5	12	2	2	5
100	6	98	82	82	87
100	7	23	7	8	14
100	8	100	99	99	100
100	9	12	0	0	2
100	10	12	0	0	1
100	11	9	1	1	6
100	12	13	2	2	4
100	13	28	3	3	4
100	14	14	6	8	9
100	15	26	2	2	8
100	16	33	6	7	16
100	17	21	3	6	8
100	18	20	2	2	4
100	19	38	6	7	15
100	20	17	2	2	5

Condition 10 (20 items test length, 1000/1000, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	6	0	0	3
100	2	39	10	13	28
100	3	93	71	75	91
100	4	12	2	2	6
100	5	34	4	5	14
100	6	100	100	100	100
100	7	49	11	14	32
100	8	100	100	100	100
100	9	24	4	5	11
100	10	19	1	1	6
100	11	14	6	6	9
100	12	23	3	4	16
100	13	45	10	12	30
100	14	44	7	10	27
100	15	36	10	12	25
100	16	61	15	18	42
100	17	24	7	7	14
100	18	29	8	8	18
100	19	55	19	19	36
100	20	42	11	13	24

Condition 11 (20 items test length, 700/300, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	7	0	0	1
100	2	15	3	3	8
100	3	56	15	15	25
100	4	6	0	0	1
100	5	21	1	1	8
100	6	95	76	76	82
100	7	29	7	7	12
100	8	100	93	93	98
100	9	12	0	0	3
100	10	10	2	2	5
100	11	8	1	1	2
100	12	10	3	3	4
100	13	20	3	3	10
100	14	19	2	2	7
100	15	31	3	3	7
100	16	33	3	4	14
100	17	18	1	2	2
100	18	16	4	5	10
100	19	21	4	4	8
100	20	24	3	3	8

Condition 12 (20 items test length, 1500/500, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	10	0	0	3
100	2	16	5	5	10
100	3	78	34	35	45
100	4	7	0	0	1
100	5	16	2	2	6
100	6	92	62	62	74
100	7	21	1	3	5
100	8	99	97	97	98
100	9	11	1	1	3
100	10	4	1	1	3
100	11	13	0	0	1
100	12	9	1	1	1
100	13	17	3	3	5
100	14	18	3	3	7
100	15	14	1	1	3
100	16	27	7	8	11
100	17	12	1	2	4
100	18	12	0	0	2
100	19	30	6	6	10
100	20	25	3	3	5

Condition 13 (40 items test length, 500/500, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	12	0	0	2
100	2	8	0	0	0
100	3	5	0	0	1
100	4	9	0	0	3
100	5	100	98	98	99
100	6	3	0	0	1
100	7	8	0	0	0
100	8	4	0	0	2
100	9	12	0	0	1
100	10	40	10	10	17
100	11	11	0	0	1
100	12	3	0	0	0
100	13	8	0	0	1
100	14	3	0	0	0
100	15	99	93	93	97
100	16	5	0	0	0
100	17	1	2	2	2
100	18	16	0	0	2
100	19	11	0	0	0
100	20	16	0	0	3
100	21	11	0	0	1
100	22	9	0	0	0
100	23	10	4	4	5
100	24	8	0	0	1
100	25	78	24	26	44
100	26	6	0	0	0
100	27	7	0	1	2
100	28	7	1	1	2
100	29	5	0	0	0
100	30	6	0	0	0
100	31	7	1	1	2
100	32	6	0	0	1
100	33	8	1	1	1
100	34	12	1	1	2
100	35	74	30	31	46
100	36	8	0	0	1
100	37	6	0	0	0
100	38	10	0	0	1
100	39	16	0	0	3
100	40	12	0	0	1

Condition 14 (40 items test length, 1000/1000, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	9	0	0	3
100	2	11	0	0	3
100	3	5	0	0	0
100	4	10	0	0	2
100	5	100	100	100	100
100	6	11	0	1	1
100	7	11	0	0	4
100	8	12	0	0	2
100	9	11	2	2	2
100	10	70	21	22	38
100	11	7	0	0	2
100	12	7	1	1	2
100	13	7	0	0	1
100	14	5	0	0	1
100	15	100	100	100	100
100	16	8	0	0	1
100	17	12	1	1	2
100	18	16	1	1	4
100	19	10	0	0	2
100	20	26	2	2	7
100	21	11	0	0	5
100	22	13	1	1	2
100	23	9	1	1	3
100	24	13	0	0	4
100	25	99	73	73	86
100	26	16	0	0	3
100	27	13	0	1	1
100	28	11	2	2	4
100	29	12	0	0	4
100	30	11	1	1	4
100	31	7	0	0	0
100	32	6	0	0	1
100	33	17	0	1	6
100	34	19	1	1	4
100	35	95	60	61	82
100	36	14	1	1	3
100	37	16	2	2	8
100	38	18	2	2	7
100	39	28	2	2	10
100	40	18	3	3	6

Condition 15 (40 items test length, 700/300, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	5	0	0	1
100	2	7	0	0	0
100	3	4	0	0	0
100	4	10	1	1	1
100	5	96	72	72	80
100	6	3	0	0	0
100	7	6	1	1	1
100	8	2	0	0	0
100	9	6	0	0	0
100	10	22	4	4	5
100	11	6	0	0	0
100	12	4	0	0	0
100	13	8	1	1	1
100	14	7	0	0	0
100	15	94	58	59	67
100	16	3	0	0	0
100	17	8	0	0	0
100	18	7	1	1	1
100	19	8	0	0	0
100	20	24	0	0	3
100	21	13	0	0	0
100	22	9	0	0	1
100	23	10	0	0	0
100	24	6	0	0	0
100	25	56	16	17	25
100	26	5	0	0	1
100	27	6	1	1	1
100	28	4	1	1	1
100	29	7	1	1	1
100	30	7	0	0	0
100	31	3	0	0	0
100	32	7	0	0	1
100	33	5	1	1	1
100	34	13	1	1	4
100	35	48	10	10	16
100	36	12	0	0	1
100	37	3	0	0	0
100	38	8	1	1	2
100	39	9	1	1	2
100	40	12	0	0	1

Condition 16 (40 items test length, 1500/500, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	5	0	0	0
100	2	2	0	0	0
100	3	9	1	1	2
100	4	6	0	0	1
100	5	100	96	96	97
100	6	3	0	0	0
100	7	9	0	0	1
100	8	7	0	0	0
100	9	11	0	0	1
100	10	33	4	4	10
100	11	3	0	0	0
100	12	3	0	0	0
100	13	6	0	0	0
100	14	8	0	0	1
100	15	99	91	91	92
100	16	10	0	0	1
100	17	7	1	1	2
100	18	14	1	1	4
100	19	10	1	1	4
100	20	14	1	1	2
100	21	9	0	0	0
100	22	11	1	1	2
100	23	8	0	0	0
100	24	11	0	0	4
100	25	80	35	36	52
100	26	11	0	0	1
100	27	6	0	0	1
100	28	13	0	0	3
100	29	6	0	0	0
100	30	5	0	0	1
100	31	5	0	0	1
100	32	7	0	0	2
100	33	17	0	0	3
100	34	3	0	0	0
100	35	74	23	24	37
100	36	11	0	0	2
100	37	9	0	0	1
100	38	16	0	0	6
100	39	8	0	0	2
100	40	10	2	2	4

Condition 17 (40 items test length, 500/500, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	7	0	0	1
100	2	9	1	1	2
100	3	6	1	1	1
100	4	9	0	0	3
100	5	100	99	99	99
100	6	3	0	0	0
100	7	6	1	1	1
100	8	9	0	0	2
100	9	13	0	0	3
100	10	40	4	4	11
100	11	2	0	0	0
100	12	5	0	0	0
100	13	10	1	1	1
100	14	8	0	1	2
100	15	100	89	89	96
100	16	6	0	0	0
100	17	8	1	1	2
100	18	14	0	0	0
100	19	7	0	0	1
100	20	26	3	4	8
100	21	8	0	0	0
100	22	5	0	0	0
100	23	10	0	0	0
100	24	8	1	1	1
100	25	70	15	15	28
100	26	10	0	0	1
100	27	13	0	0	2
100	28	7	0	0	0
100	29	4	0	0	0
100	30	4	0	0	0
100	31	6	1	1	1
100	32	3	0	0	1
100	33	11	0	0	4
100	34	5	0	0	0
100	35	68	22	22	33
100	36	7	1	1	2
100	37	10	1	1	2
100	38	11	0	0	2
100	39	7	0	0	3
100	40	15	0	0	2

Condition 18 (40 items test length, 1000/1000, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	10	0	0	1
100	2	13	0	0	0
100	3	3	0	0	0
100	4	8	0	0	0
100	5	100	100	100	100
100	6	7	0	0	1
100	7	13	0	0	4
100	8	21	1	1	2
100	9	6	0	0	2
100	10	61	20	21	32
100	11	9	1	1	2
100	12	3	0	0	2
100	13	4	0	0	1
100	14	15	0	0	2
100	15	100	100	100	100
100	16	6	0	0	2
100	17	15	1	1	3
100	18	13	1	1	3
100	19	11	0	0	2
100	20	36	6	8	15
100	21	8	0	0	1
100	22	15	0	0	3
100	23	10	0	0	2
100	24	8	0	0	2
100	25	95	66	66	81
100	26	16	3	3	5
100	27	12	0	0	2
100	28	10	0	0	3
100	29	11	2	2	4
100	30	10	0	0	2
100	31	2	1	1	1
100	32	5	0	0	1
100	33	11	0	0	4
100	34	14	1	1	3
100	35	94	61	62	77
100	36	11	1	1	3
100	37	19	2	2	6
100	38	11	1	1	3
100	39	22	0	1	6
100	40	18	0	1	6

Condition 19 (40 items test length, 700/300, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	4	0	0	0
100	2	8	0	0	0
100	3	7	0	0	0
100	4	3	0	0	0
100	5	99	75	75	77
100	6	4	0	0	0
100	7	8	0	0	1
100	8	5	0	0	0
100	9	8	1	1	1
100	10	22	1	1	2
100	11	11	1	1	2
100	12	4	0	0	0
100	13	6	1	1	2
100	14	4	1	1	1
100	15	94	46	47	60
100	16	7	1	1	1
100	17	10	0	0	1
100	18	7	0	0	0
100	19	9	0	0	0
100	20	24	3	3	4
100	21	11	1	1	2
100	22	6	0	0	0
100	23	6	0	0	1
100	24	6	0	0	0
100	25	54	11	12	19
100	26	7	0	0	1
100	27	10	0	0	0
100	28	7	0	0	0
100	29	4	0	0	0
100	30	6	0	0	1
100	31	4	0	0	1
100	32	6	1	1	3
100	33	13	0	0	1
100	34	12	1	1	1
100	35	44	8	8	11
100	36	8	0	0	2
100	37	4	0	0	1
100	38	6	0	0	0
100	39	13	0	0	0
100	40	13	1	1	1

Condition 20 (40 items test length, 1500/500, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	6	0	0	3
100	2	6	0	0	2
100	3	8	0	0	0
100	4	7	0	0	1
100	5	100	91	92	98
100	6	5	0	0	1
100	7	4	0	0	0
100	8	8	0	0	2
100	9	9	0	0	1
100	10	36	4	4	13
100	11	5	0	0	2
100	12	3	1	1	1
100	13	7	0	0	1
100	14	10	1	1	1
100	15	99	90	90	94
100	16	6	0	0	1
100	17	8	1	1	1
100	18	10	2	2	2
100	19	8	0	0	2
100	20	21	1	1	5
100	21	10	1	1	1
100	22	5	0	0	0
100	23	4	1	1	2
100	24	9	0	0	1
100	25	76	25	25	40
100	26	7	1	1	2
100	27	5	0	0	1
100	28	9	1	1	2
100	29	5	0	0	0
100	30	3	0	0	0
100	31	5	0	0	1
100	32	7	0	0	0
100	33	10	0	0	4
100	34	3	1	1	1
100	35	66	28	28	39
100	36	13	1	1	4
100	37	6	0	0	2
100	38	9	1	2	3
100	39	10	0	0	5
100	40	17	1	1	2

Condition 21 (40 items test length, 500/500, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	6	0	0	0
100	2	1	0	0	1
100	3	13	1	1	3
100	4	22	3	3	5
100	5	86	38	40	58
100	6	5	0	0	0
100	7	3	0	0	0
100	8	2	0	0	1
100	9	12	0	0	0
100	10	10	0	0	0
100	11	89	51	51	64
100	12	96	52	53	68
100	13	8	0	0	0
100	14	16	2	2	4
100	15	91	48	49	68
100	16	24	0	0	7
100	17	11	0	0	0
100	18	10	1	1	4
100	19	6	0	0	0
100	20	8	0	0	0
100	21	5	0	0	0
100	22	6	0	0	1
100	23	9	0	0	1
100	24	5	0	0	0
100	25	14	0	0	0
100	26	12	0	0	2
100	27	12	0	0	2
100	28	11	0	0	2
100	29	11	0	0	1
100	30	5	1	1	1
100	31	16	2	2	5
100	32	28	2	2	3
100	33	5	0	0	0
100	34	5	0	0	0
100	35	15	0	0	2
100	36	9	0	0	1
100	37	23	0	0	4
100	38	28	2	2	4
100	39	12	0	0	0
100	40	16	2	2	4

Condition 22 (40 items test length, 1000/1000, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	4	1	1	1
100	2	4	1	1	1
100	3	25	5	5	13
100	4	26	3	3	9
100	5	96	68	69	88
100	6	7	0	0	0
100	7	4	0	0	0
100	8	8	0	0	3
100	9	10	0	0	3
100	10	8	0	0	1
100	11	99	85	86	95
100	12	100	92	92	96
100	13	36	7	7	18
100	14	40	7	7	17
100	15	100	88	88	96
100	16	37	3	4	15
100	17	6	0	0	1
100	18	12	0	0	5
100	19	14	0	0	4
100	20	24	2	2	5
100	21	10	2	2	4
100	22	13	0	0	1
100	23	6	0	0	3
100	24	8	2	2	2
100	25	30	1	1	7
100	26	26	1	1	7
100	27	19	2	2	4
100	28	21	3	4	6
100	29	23	2	2	7
100	30	9	0	0	1
100	31	36	10	10	17
100	32	46	9	10	21
100	33	25	0	0	8
100	34	18	2	2	5
100	35	29	3	3	11
100	36	22	1	1	7
100	37	32	8	8	15
100	38	40	7	7	15
100	39	34	3	3	8
100	40	24	2	2	10

Condition 23 (40 items test length, 700/300, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	6	0	0	0
100	2	5	0	0	1
100	3	8	1	1	1
100	4	13	1	1	1
100	5	62	15	15	21
100	6	6	0	0	0
100	7	6	0	0	0
100	8	3	0	0	0
100	9	5	1	1	3
100	10	8	0	0	0
100	11	75	30	30	36
100	12	70	28	28	34
100	13	14	1	1	1
100	14	11	1	1	1
100	15	71	27	28	31
100	16	18	2	2	2
100	17	11	0	0	1
100	18	5	0	0	0
100	19	10	2	2	3
100	20	13	0	0	2
100	21	5	0	0	0
100	22	5	0	0	0
100	23	5	0	0	0
100	24	6	0	0	0
100	25	13	0	0	0
100	26	7	0	0	2
100	27	12	1	1	1
100	28	7	1	1	1
100	29	13	1	1	1
100	30	9	1	1	1
100	31	15	0	0	0
100	32	13	1	1	3
100	33	11	0	0	1
100	34	7	2	2	4
100	35	10	0	0	1
100	36	7	0	0	0
100	37	12	0	0	1
100	38	10	1	1	1
100	39	13	0	0	0
100	40	10	1	1	1

Condition 24(40 items test length, 1500/500, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	7	0	0	0
100	2	5	0	0	0
100	3	15	1	1	2
100	4	13	2	2	4
100	5	70	31	33	47
100	6	2	0	0	0
100	7	1	0	0	0
100	8	4	0	0	1
100	9	15	0	0	3
100	10	2	0	0	1
100	11	89	55	55	68
100	12	92	50	51	62
100	13	19	1	1	3
100	14	17	1	1	4
100	15	88	55	55	68
100	16	18	3	3	5
100	17	8	1	1	2
100	18	10	0	0	0
100	19	9	0	0	0
100	20	14	1	1	2
100	21	11	0	0	2
100	22	3	1	1	1
100	23	7	0	0	1
100	24	12	1	1	3
100	25	13	1	1	4
100	26	17	2	2	5
100	27	16	3	3	4
100	28	13	2	2	3
100	29	7	1	1	1
100	30	6	0	0	1
100	31	28	3	3	7
100	32	21	3	3	4
100	33	12	2	2	4
100	34	4	0	0	2
100	35	15	1	1	5
100	36	9	0	0	1
100	37	18	2	2	3
100	38	24	2	2	10
100	39	16	2	2	4
100	40	15	2	2	3

APPENDIX E

DETECTION OF FALSE POSITIVES IN LOGISTIC REGRESSION PROCEDURE BEFORE AND AFTER THREE ADJUSTMENTS

Condition 25(20 items test length, 500/500, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	9	0	0	1
100	2	14	1	2	6
100	3	100	100	100	100
100	4	13	2	2	5
100	5	4	0	1	1
100	6	5	0	0	1
100	7	21	2	2	5
100	8	100	100	100	100
100	9	18	4	4	7
100	10	12	1	1	6
100	11	15	0	0	4
100	12	13	3	3	8
100	13	6	1	1	1
100	14	12	0	0	2
100	15	9	0	0	1
100	16	16	3	3	5
100	17	20	1	2	8
100	18	19	4	4	8
100	19	14	2	2	3
100	20	14	2	2	2

Condition 26(20 items test length, 1000/1000, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	19	2	2	7
100	2	14	1	2	5
100	3	100	100	100	100
100	4	15	2	2	5
100	5	11	0	1	5
100	6	13	3	4	8
100	7	18	4	4	10
100	8	100	100	100	100
100	9	22	3	3	8
100	10	19	1	1	6
100	11	27	4	4	14
100	12	28	5	5	14
100	13	20	1	1	5
100	14	10	1	1	5
100	15	10	0	0	1
100	16	25	2	3	12
100	17	36	6	6	14
100	18	27	6	6	14
100	19	19	2	3	5
100	20	24	3	4	10

Condition 27(20 items test length, 700/300, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	18	0	1	3
100	2	10	1	1	3
100	3	100	99	99	100
100	4	8	1	1	1
100	5	8	0	0	0
100	6	7	1	1	1
100	7	10	1	1	3
100	8	100	100	100	100
100	9	13	3	4	5
100	10	15	2	2	3
100	11	12	2	2	4
100	12	7	1	2	2
100	13	8	0	0	4
100	14	8	0	0	2
100	15	7	0	0	0
100	16	14	2	2	3
100	17	12	1	1	4
100	18	20	3	4	9
100	19	9	0	0	2
100	20	24	0	0	5

Condition 28(20 items test length, 1500/500, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	7	2	2	3
100	2	8	0	0	2
100	3	99	90	90	91
100	4	14	0	0	0
100	5	11	0	0	3
100	6	5	1	1	1
100	7	6	1	1	1
100	8	100	99	99	99
100	9	10	1	1	4
100	10	10	0	0	1
100	11	17	3	3	5
100	12	10	0	0	1
100	13	10	0	0	2
100	14	9	0	0	4
100	15	7	0	0	0
100	16	8	1	1	3
100	17	10	0	0	2
100	18	8	1	1	1
100	19	8	1	1	1
100	20	14	1	1	3

Condition 29(20 items test length, 500/500, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	8	1	1	2
100	2	11	0	0	2
100	3	100	100	100	100
100	4	11	0	0	1
100	5	5	0	0	0
100	6	8	1	1	1
100	7	15	2	2	4
100	8	100	100	100	100
100	9	12	3	3	4
100	10	12	0	0	2
100	11	18	4	4	8
100	12	14	1	1	5
100	13	13	0	0	1
100	14	12	1	1	3
100	15	7	0	0	0
100	16	12	0	0	5
100	17	9	1	1	2
100	18	19	3	3	7
100	19	11	0	0	3
100	20	17	5	5	6

Condition 30(20 items test length, 1000/1000, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	17	2	2	4
100	2	17	3	3	6
100	3	100	100	100	100
100	4	18	2	3	5
100	5	19	1	1	1
100	6	14	1	2	4
100	7	16	4	4	10
100	8	100	100	100	100
100	9	22	4	4	7
100	10	16	1	1	5
100	11	18	3	4	6
100	12	24	3	3	11
100	13	30	6	6	11
100	14	9	2	2	2
100	15	7	0	0	1
100	16	19	1	1	4
100	17	24	6	7	11
100	18	24	5	5	9
100	19	13	1	1	5
100	20	25	4	4	11

Condition 31(20 items test length, 700/300, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	6	1	1	1
100	2	13	1	1	2
100	3	99	95	95	97
100	4	11	0	0	3
100	5	10	2	2	3
100	6	4	0	0	0
100	7	11	2	3	5
100	8	100	100	100	100
100	9	15	0	1	4
100	10	16	1	1	3
100	11	6	0	0	1
100	12	5	0	1	1
100	13	6	0	0	0
100	14	12	2	2	3
100	15	5	0	0	1
100	16	9	0	0	1
100	17	7	1	1	1
100	18	14	1	1	7
100	19	11	3	3	5
100	20	14	0	0	6

Condition 32(20 items test length, 1500/500, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	10	1	1	1
100	2	7	0	0	0
100	3	100	94	94	97
100	4	7	1	1	2
100	5	10	0	0	2
100	6	9	0	0	2
100	7	9	0	0	2
100	8	100	100	100	100
100	9	8	0	0	3
100	10	9	2	2	4
100	11	18	4	4	5
100	12	14	0	0	5
100	13	10	1	1	2
100	14	6	0	0	0
100	15	13	0	0	2
100	16	10	1	1	2
100	17	10	0	1	1
100	18	11	1	1	2
100	19	14	1	1	3
100	20	19	0	0	2

Condition 33(20 items test length, 500/500, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	34	6	9	27
100	2	8	1	1	4
100	3	24	2	5	18
100	4	58	21	30	52
100	5	100	98	99	100
100	6	96	79	83	93
100	7	21	4	4	15
100	8	100	98	99	100
100	9	67	33	39	65
100	10	74	29	40	69
100	11	76	30	39	71
100	12	69	31	36	64
100	13	100	100	100	100
100	14	100	100	100	100
100	15	100	100	100	100
100	16	20	2	4	13
100	17	84	52	58	79
100	18	88	53	61	85
100	19	100	100	100	100
100	20	88	61	63	85

Condition 34(20 items test length, 1000/1000, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	68	31	45	64
100	2	24	2	7	22
100	3	32	8	17	30
100	4	81	46	62	79
100	5	100	100	100	100
100	6	100	100	100	100
100	7	31	6	10	29
100	8	100	100	100	100
100	9	93	69	79	93
100	10	95	72	83	95
100	11	97	70	86	97
100	12	96	75	87	95
100	13	100	100	100	100
100	14	100	100	100	100
100	15	100	100	100	100
100	16	34	9	15	33
100	17	99	88	93	98
100	18	100	90	94	100
100	19	100	100	100	100
100	20	99	95	97	99

Condition 35(20 items test length, 700/300, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	22	7	8	20
100	2	9	2	3	7
100	3	13	3	3	11
100	4	29	5	10	23
100	5	94	78	80	93
100	6	96	72	77	93
100	7	16	4	6	13
100	8	99	91	94	99
100	9	58	16	19	45
100	10	55	17	22	50
100	11	64	17	26	59
100	12	56	13	20	48
100	13	100	96	97	100
100	14	100	96	98	100
100	15	100	100	100	100
100	16	25	0	1	16
100	17	60	30	32	54
100	18	71	27	31	67
100	19	100	99	99	100
100	20	81	43	50	75

Condition 36(20 items test length, 1500/500, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	40	6	9	36
100	2	13	3	4	10
100	3	20	8	9	16
100	4	48	10	19	38
100	5	98	81	84	97
100	6	89	57	64	88
100	7	11	0	0	9
100	8	99	88	93	99
100	9	63	26	32	58
100	10	57	18	24	48
100	11	74	32	40	69
100	12	64	22	31	58
100	13	99	98	98	99
100	14	100	98	99	100
100	15	100	100	100	100
100	16	21	4	4	17
100	17	73	27	31	67
100	18	74	27	34	67
100	19	100	99	99	100
100	20	87	45	51	81

Condition 37(40 items test length, 500/500, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	13	0	0	2
100	2	11	0	0	0
100	3	6	0	0	0
100	4	8	0	0	2
100	5	100	99	99	100
100	6	4	0	0	1
100	7	8	0	0	0
100	8	5	0	0	2
100	9	13	1	1	4
100	10	52	12	14	22
100	11	9	0	0	0
100	12	4	0	0	0
100	13	9	0	0	1
100	14	4	0	0	0
100	15	100	96	97	98
100	16	3	0	0	0
100	17	8	1	1	5
100	18	12	0	0	3
100	19	12	0	0	0
100	20	7	0	0	0
100	21	12	1	1	1
100	22	9	0	0	1
100	23	12	5	5	6
100	24	7	0	0	0
100	25	95	73	74	84
100	26	4	0	0	0
100	27	8	0	0	1
100	28	7	2	2	2
100	29	8	0	0	1
100	30	15	0	0	3
100	31	5	0	0	2
100	32	6	0	0	1
100	33	10	1	1	1
100	34	9	0	0	2
100	35	86	49	49	60
100	36	5	0	0	1
100	37	5	0	0	2
100	38	14	0	0	2
100	39	13	0	0	2
100	40	7	0	0	1

Condition 38(40 items test length, 1000/1000, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	11	0	0	6
100	2	12	0	0	2
100	3	6	0	0	0
100	4	10	0	0	3
100	5	100	100	100	100
100	6	12	1	1	4
100	7	14	1	1	5
100	8	12	1	1	4
100	9	12	2	2	3
100	10	88	36	36	59
100	11	7	1	1	2
100	12	7	1	1	2
100	13	7	0	0	1
100	14	4	0	0	1
100	15	100	100	100	100
100	16	6	0	0	1
100	17	11	1	1	3
100	18	17	1	1	6
100	19	14	1	1	2
100	20	9	1	1	2
100	21	17	1	1	6
100	22	13	1	2	3
100	23	11	1	1	4
100	24	15	0	0	6
100	25	100	100	100	100
100	26	18	0	0	7
100	27	11	0	0	3
100	28	14	2	2	3
100	29	12	0	1	4
100	30	33	6	6	9
100	31	7	0	0	0
100	32	7	0	0	1
100	33	17	0	0	8
100	34	18	1	1	4
100	35	98	87	87	95
100	36	12	0	0	3
100	37	16	4	4	8
100	38	17	2	2	8
100	39	27	3	3	7
100	40	15	3	3	7

Condition 39(40 items test length, 700/300, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	3	0	0	1
100	2	7	1	1	1
100	3	3	0	0	0
100	4	8	1	1	1
100	5	98	82	82	88
100	6	4	0	0	0
100	7	3	1	1	1
100	8	3	0	0	0
100	9	11	0	0	0
100	10	27	2	2	7
100	11	4	0	0	0
100	12	4	0	0	0
100	13	7	1	1	1
100	14	7	0	0	0
100	15	95	68	68	78
100	16	6	0	0	0
100	17	5	0	0	0
100	18	4	1	1	1
100	19	10	0	0	1
100	20	14	0	0	1
100	21	13	0	0	3
100	22	9	0	0	0
100	23	6	0	0	0
100	24	6	0	0	0
100	25	84	40	40	53
100	26	5	0	0	0
100	27	10	0	0	0
100	28	3	0	0	1
100	29	9	0	0	1
100	30	11	0	0	1
100	31	5	0	0	0
100	32	7	0	0	1
100	33	4	1	1	1
100	34	17	1	1	3
100	35	67	19	19	26
100	36	11	0	0	2
100	37	7	0	0	0
100	38	10	0	0	3
100	39	12	1	1	3
100	40	13	0	0	1

Condition 40(40 items test length, 1500/500, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	8	0	0	0
100	2	4	0	0	0
100	3	9	0	0	2
100	4	6	0	0	1
100	5	100	98	99	100
100	6	5	0	0	0
100	7	9	0	0	1
100	8	6	0	0	1
100	9	12	0	0	2
100	10	48	8	8	15
100	11	3	0	0	0
100	12	3	0	0	0
100	13	6	0	0	1
100	14	9	0	0	1
100	15	100	93	93	98
100	16	9	0	0	1
100	17	9	0	0	1
100	18	11	0	0	4
100	19	13	2	2	3
100	20	8	0	0	1
100	21	9	0	0	0
100	22	12	1	1	1
100	23	7	0	0	0
100	24	9	0	0	1
100	25	96	73	73	83
100	26	11	0	0	1
100	27	6	0	0	1
100	28	10	2	3	3
100	29	7	0	0	2
100	30	12	2	2	3
100	31	4	0	0	1
100	32	7	0	0	2
100	33	16	0	0	3
100	34	4	0	0	0
100	35	83	41	44	63
100	36	10	0	0	1
100	37	8	0	0	2
100	38	14	1	1	9
100	39	7	0	1	3
100	40	10	1	1	4

Condition 41(40 items test length, 500/500, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	9	0	0	3
100	2	11	0	1	3
100	3	8	1	1	1
100	4	9	0	0	3
100	5	100	99	99	100
100	6	4	0	0	0
100	7	9	1	1	2
100	8	14	0	0	2
100	9	14	1	1	4
100	10	51	6	6	20
100	11	2	0	0	0
100	12	5	0	0	1
100	13	11	0	0	1
100	14	8	0	0	1
100	15	100	96	96	99
100	16	4	0	0	0
100	17	9	1	1	2
100	18	13	0	0	1
100	19	10	0	0	1
100	20	14	1	1	3
100	21	10	0	0	0
100	22	5	0	0	1
100	23	9	0	0	3
100	24	10	1	1	1
100	25	93	54	54	73
100	26	8	1	1	2
100	27	17	0	0	2
100	28	9	0	0	0
100	29	5	0	0	1
100	30	13	0	0	0
100	31	6	1	1	1
100	32	4	0	0	0
100	33	11	0	0	2
100	34	6	0	0	0
100	35	85	32	33	56
100	36	7	1	1	1
100	37	10	0	0	3
100	38	9	1	1	2
100	39	8	1	1	3
100	40	13	0	0	2

Condition 42(40 items test length, 1000/1000, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	11	0	0	1
100	2	13	0	0	1
100	3	2	0	0	0
100	4	12	0	0	2
100	5	100	100	100	100
100	6	9	0	0	4
100	7	17	1	1	5
100	8	18	1	1	4
100	9	9	1	1	1
100	10	76	32	33	47
100	11	8	0	0	4
100	12	3	0	0	1
100	13	6	1	1	1
100	14	12	0	0	2
100	15	100	100	100	100
100	16	6	0	0	1
100	17	14	1	1	6
100	18	19	0	0	6
100	19	12	0	0	2
100	20	19	1	2	8
100	21	10	1	1	2
100	22	15	0	0	3
100	23	13	0	0	1
100	24	11	1	2	2
100	25	99	96	96	98
100	26	15	1	1	4
100	27	12	2	2	4
100	28	12	0	0	2
100	29	10	1	1	3
100	30	23	1	1	7
100	31	2	1	1	1
100	32	6	0	0	1
100	33	13	0	0	4
100	34	15	0	1	2
100	35	99	78	78	92
100	36	14	1	1	2
100	37	20	1	1	5
100	38	12	1	1	4
100	39	19	1	1	5
100	40	16	0	0	6

Condition 43(40 items test length, 700/300, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	3	0	0	0
100	2	11	0	0	1
100	3	6	0	0	0
100	4	8	0	0	0
100	5	100	84	84	86
100	6	5	0	0	0
100	7	9	0	0	2
100	8	8	0	0	0
100	9	10	1	1	1
100	10	26	2	2	5
100	11	10	1	1	2
100	12	6	0	0	0
100	13	5	1	1	1
100	14	3	1	1	1
100	15	95	57	58	66
100	16	4	1	1	1
100	17	7	0	0	0
100	18	5	1	1	1
100	19	6	0	0	1
100	20	16	1	1	4
100	21	12	2	2	2
100	22	4	0	0	0
100	23	6	0	0	0
100	24	8	0	0	0
100	25	83	27	27	41
100	26	8	0	0	1
100	27	7	0	0	0
100	28	6	0	0	0
100	29	5	0	0	0
100	30	10	0	0	1
100	31	3	0	0	0
100	32	7	1	1	3
100	33	8	0	0	1
100	34	12	1	1	1
100	35	60	13	13	20
100	36	10	0	0	2
100	37	5	0	0	1
100	38	5	0	0	0
100	39	12	0	0	1
100	40	12	1	1	3

Condition 44(40 items test length, 1500/500, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	9	0	0	2
100	2	6	0	0	1
100	3	7	0	0	0
100	4	6	0	0	0
100	5	100	100	100	100
100	6	6	0	0	0
100	7	6	0	0	0
100	8	9	0	0	1
100	9	9	0	0	2
100	10	47	6	6	14
100	11	4	0	0	2
100	12	4	0	0	1
100	13	9	0	0	1
100	14	11	1	1	1
100	15	100	93	93	97
100	16	7	0	0	1
100	17	10	1	1	3
100	18	10	0	0	2
100	19	5	0	0	3
100	20	15	0	0	2
100	21	8	1	1	2
100	22	7	0	0	1
100	23	5	0	0	2
100	24	11	0	0	1
100	25	94	69	69	81
100	26	10	0	0	3
100	27	6	0	0	2
100	28	14	1	1	3
100	29	4	0	0	1
100	30	11	0	0	1
100	31	5	0	0	2
100	32	8	0	0	0
100	33	11	0	0	5
100	34	3	1	1	1
100	35	81	43	43	55
100	36	13	3	3	4
100	37	9	0	0	1
100	38	9	0	0	3
100	39	9	1	1	3
100	40	13	1	1	2

Condition 45(40 items test length, 500/500, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	39	5	8	32
100	2	36	4	5	28
100	3	13	1	1	10
100	4	13	1	3	9
100	5	24	1	2	17
100	6	50	8	9	44
100	7	61	8	12	51
100	8	53	6	8	41
100	9	99	91	93	99
100	10	86	48	54	80
100	11	93	42	48	88
100	12	94	50	56	91
100	13	1	0	0	1
100	14	10	1	1	8
100	15	96	70	71	95
100	16	14	0	0	10
100	17	75	27	30	60
100	18	73	27	34	68
100	19	77	27	31	69
100	20	45	5	7	39
100	21	74	21	27	67
100	22	73	19	22	62
100	23	66	23	30	57
100	24	72	25	29	69
100	25	77	35	39	71
100	26	100	99	99	100
100	27	100	99	99	100
100	28	100	99	100	100
100	29	100	100	100	100
100	30	100	100	100	100
100	31	7	0	0	5
100	32	8	0	1	5
100	33	83	30	35	75
100	34	87	42	53	81
100	35	10	0	0	6
100	36	84	33	38	77
100	37	100	100	100	100
100	38	100	100	100	100
100	39	88	54	57	85
100	40	91	56	60	85

Condition 46(40 items test length, 1000/1000, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	68	15	25	63
100	2	61	19	27	59
100	3	14	0	1	11
100	4	13	0	0	11
100	5	24	1	5	21
100	6	75	32	43	73
100	7	83	32	39	77
100	8	82	41	50	81
100	9	100	100	100	100
100	10	98	90	94	98
100	11	99	90	93	99
100	12	100	91	92	100
100	13	20	1	5	19
100	14	22	1	5	20
100	15	100	94	98	100
100	16	22	1	2	19
100	17	91	58	69	90
100	18	96	58	75	95
100	19	98	61	72	98
100	20	80	34	49	76
100	21	95	68	78	93
100	22	100	59	73	97
100	23	95	56	72	92
100	24	97	68	76	95
100	25	98	75	90	98
100	26	100	100	100	100
100	27	100	100	100	100
100	28	100	100	100	100
100	29	100	100	100	100
100	30	100	100	100	100
100	31	17	1	4	14
100	32	20	1	3	17
100	33	99	89	93	99
100	34	99	87	94	99
100	35	30	5	7	26
100	36	98	86	92	97
100	37	100	100	100	100
100	38	100	100	100	100
100	39	100	96	97	100
100	40	99	94	97	99

Condition 47(40 items test length, 700/300, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	21	0	1	13
100	2	20	0	0	7
100	3	4	0	0	3
100	4	7	1	1	2
100	5	17	1	2	12
100	6	35	6	7	25
100	7	27	6	8	17
100	8	33	4	4	23
100	9	93	59	61	90
100	10	68	24	30	57
100	11	69	20	24	55
100	12	68	23	25	53
100	13	11	1	1	7
100	14	4	0	0	1
100	15	74	32	38	70
100	16	11	0	0	4
100	17	55	5	9	38
100	18	50	8	8	38
100	19	56	12	13	38
100	20	49	6	7	32
100	21	51	13	13	33
100	22	44	7	8	30
100	23	43	5	5	26
100	24	47	6	8	34
100	25	57	13	14	47
100	26	100	81	84	98
100	27	98	83	87	97
100	28	97	85	87	94
100	29	100	97	97	100
100	30	99	88	88	96
100	31	8	0	0	2
100	32	9	0	0	3
100	33	62	14	17	49
100	34	68	12	17	60
100	35	9	0	0	7
100	36	59	13	13	44
100	37	100	92	95	100
100	38	100	96	97	100
100	39	75	21	26	65
100	40	63	18	24	57

Condition 48(40 items test length, 1500/500, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	35	5	6	28
100	2	37	7	7	27
100	3	10	0	0	8
100	4	9	1	1	6
100	5	13	0	0	10
100	6	50	2	4	40
100	7	46	6	7	37
100	8	51	11	14	44
100	9	98	91	92	98
100	10	92	45	50	90
100	11	89	55	59	89
100	12	89	50	56	79
100	13	8	0	1	8
100	14	7	1	1	4
100	15	95	71	74	93
100	16	11	0	2	8
100	17	71	24	30	63
100	18	77	24	30	64
100	19	69	22	25	62
100	20	56	4	8	44
100	21	71	16	21	60
100	22	72	18	20	61
100	23	69	16	22	59
100	24	74	25	30	66
100	25	77	32	37	69
100	26	100	100	100	100
100	27	100	100	100	100
100	28	100	100	100	100
100	29	100	100	100	100
100	30	100	98	98	100
100	31	12	0	2	8
100	32	10	0	0	9
100	33	89	39	45	84
100	34	84	31	37	76
100	35	11	0	1	6
100	36	90	41	45	85
100	37	100	100	100	100
100	38	100	100	100	100
100	39	88	53	60	86
100	40	90	43	47	85

APPENDIX F

DETECTION OF FALSE POSITIVES IN THE DFIT METHOD BEFORE AND AFTER THREE ADJUSTMENTS

Condition 49(20 items test length, 1000/1000, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	10	0	0	2
100	2	2	1	1	1
100	3	90	29	20	38
100	4	5	0	0	0
100	5	1	0	0	0
100	6	7	2	2	2
100	7	7	2	2	3
100	8	100	100	75	100
100	9	6	1	0	2
100	10	7	0	0	0
100	11	7	0	0	0
100	12	6	0	0	0
100	13	0	0	0	0
100	14	2	0	0	0
100	15	0	0	0	0
100	16	9	5	0	7
100	17	8	0	0	0
100	18	10	0	0	3
100	19	4	1	0	1
100	20	10	1	0	3

Condition 50(20 items test length, 1000/1000, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	8	1	0	1
100	2	7	1	0	1
100	3	91	37	21	41
100	4	4	1	1	2
100	5	1	0	0	0
100	6	10	2	2	3
100	7	15	3	3	6
100	8	100	100	73	100
100	9	10	0	0	0
100	10	6	0	0	0
100	11	6	1	0	1
100	12	9	1	0	2
100	13	3	0	0	2
100	14	2	1	0	1
100	15	0	0	0	0
100	16	5	4	0	4
100	17	6	0	0	0
100	18	11	2	0	4
100	19	2	0	0	0
100	20	5	1	0	1

Condition 51(20 items test length, 1000/1000, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	27	13	14	14
100	2	0	0	0	0
100	3	96	60	64	65
100	4	12	1	1	3
100	5	13	0	0	1
100	6	6	3	1	4
100	7	1	0	0	0
100	8	72	23	17	29
100	9	1	0	0	0
100	10	6	2	1	2
100	11	2	0	0	0
100	12	3	1	1	1
100	13	14	0	0	0
100	14	7	1	0	2
100	15	3	2	1	2
100	16	3	0	0	0
100	17	0	0	0	0
100	18	2	0	0	0
100	19	8	0	0	1
100	20	0	0	0	0

Condition 52(40 items test length, 1000/1000, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	10	0	0	2
100	2	7	0	0	1
100	3	1	0	0	0
100	4	4	0	0	0
100	5	97	46	46	58
100	6	9	1	1	1
100	7	8	0	0	1
100	8	11	0	0	2
100	9	7	2	2	3
100	10	26	2	2	13
100	11	3	1	1	2
100	12	5	0	0	1
100	13	3	1	1	1
100	14	10	1	1	3
100	15	100	100	100	100
100	16	2	0	0	0
100	17	8	1	1	2
100	18	12	2	2	6
100	19	7	1	1	2
100	20	71	33	33	54
100	21	9	0	0	4
100	22	6	0	0	1
100	23	12	0	0	7
100	24	16	1	1	4
100	25	86	18	18	34
100	26	12	2	2	5
100	27	12	0	0	3
100	28	11	1	1	3
100	29	3	0	0	1
100	30	0	0	0	0
100	31	3	0	0	1
100	32	2	0	0	0
100	33	13	0	0	0
100	34	12	1	1	3
100	35	99	86	86	90
100	36	14	1	1	2
100	37	13	0	0	2
100	38	7	0	0	1
100	39	12	2	2	3
100	40	34	4	4	12

Condition 53(40 items test length, 1000/1000, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	12	0	0	0
100	2	8	0	0	0
100	3	2	0	0	0
100	4	3	0	0	0
100	5	95	32	30	45
100	6	11	0	0	2
100	7	16	0	0	1
100	8	12	1	0	1
100	9	5	0	0	1
100	10	25	1	0	10
100	11	5	0	0	0
100	12	10	1	0	1
100	13	8	0	0	0
100	14	7	1	1	3
100	15	99	99	67	99
100	16	10	0	0	2
100	17	15	1	0	4
100	18	11	1	0	2
100	19	10	0	0	1
100	20	44	23	1	31
100	21	7	0	0	2
100	22	14	2	0	4
100	23	11	0	0	3
100	24	11	0	0	1
100	25	79	0	0	5
100	26	9	0	0	2
100	27	11	0	0	2
100	28	11	1	0	3
100	29	1	0	0	0
100	30	0	0	0	0
100	31	1	1	0	1
100	32	5	1	0	2
100	33	10	2	0	3
100	34	10	0	0	3
100	35	83	55	1	68
100	36	13	0	0	1
100	37	9	0	0	0
100	38	7	0	0	1
100	39	12	1	0	4
100	40	28	1	0	8

Condition 54(40 items test length, 1000/1000, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	17	8	6	9
100	2	33	12	7	16
100	3	4	1	0	1
100	4	7	0	0	1
100	5	95	46	32	64
100	6	19	4	2	6
100	7	17	5	1	7
100	8	20	2	2	4
100	9	16	1	0	5
100	10	25	2	0	5
100	11	19	0	0	5
100	12	14	3	1	7
100	13	7	0	0	1
100	14	5	0	0	0
100	15	66	10	1	23
100	16	5	0	0	0
100	17	8	0	0	2
100	18	14	4	0	5
100	19	8	2	1	3
100	20	80	19	9	33
100	21	10	1	1	3
100	22	11	3	2	4
100	23	11	2	0	2
100	24	10	2	1	2
100	25	96	76	30	82
100	26	8	2	0	3
100	27	7	0	0	1
100	28	5	0	0	1
100	29	6	1	1	1
100	30	38	10	1	16
100	31	6	0	0	1
100	32	9	4	0	4
100	33	8	0	0	0
100	34	11	1	1	1
100	35	63	6	1	13
100	36	7	0	0	0
100	37	3	0	0	0
100	38	3	0	0	0
100	39	4	0	0	0
100	40	8	1	0	1

APPENDIX G

DETECTION OF FALSE POSITIVES IN THE LORD'S CHI-SQUARE TEST BEFORE AND AFTER THREE ADJUSTMENTS

Condition 55(20 items test length, 1000/1000, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	10	0	0	2
100	2	2	1	1	1
100	3	90	29	20	38
100	4	5	0	0	0
100	5	1	0	0	0
100	6	7	2	2	2
100	7	7	2	2	3
100	8	100	100	75	100
100	9	6	1	0	2
100	10	7	0	0	0
100	11	7	0	0	0
100	12	6	0	0	0
100	13	0	0	0	0
100	14	2	0	0	0
100	15	0	0	0	0
100	16	9	5	0	7
100	17	8	0	0	0
100	18	10	0	0	3
100	19	4	1	0	1
100	20	10	1	0	3

Condition 56(20 items test length, 1000/1000, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	6	0	0	1
100	2	3	0	1	2
100	3	57	54	54	55
100	4	7	0	0	2
100	5	3	0	1	1
100	6	11	1	1	4
100	7	12	2	2	4
100	8	57	57	57	57
100	9	6	0	0	1
100	10	10	1	1	3
100	11	5	0	0	2
100	12	11	1	1	3
100	13	6	0	0	2
100	14	6	0	0	2
100	15	2	0	0	0
100	16	13	3	3	6
100	17	9	4	4	5
100	18	13	3	3	8
100	19	3	0	0	1
100	20	16	4	6	7

Condition 57(20 items test length, 1000/1000, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	27	17	17	20
100	2	15	7	7	11
100	3	43	29	30	36
100	4	32	24	24	26
100	5	30	12	14	22
100	6	9	2	2	4
100	7	33	13	13	22
100	8	53	28	30	41
100	9	36	15	17	30
100	10	38	23	23	33
100	11	29	16	16	23
100	12	36	19	21	26
100	13	28	13	13	21
100	14	24	10	11	20
100	15	18	9	9	15
100	16	53	31	34	44
100	17	46	28	29	38
100	18	47	29	29	37
100	19	39	22	22	32
100	20	58	39	41	47

Condition 58(40 items test length, 1000/1000, and no group difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	5	0	0	1
100	2	4	0	0	0
100	3	1	0	0	0
100	4	4	0	0	0
100	5	98	86	86	93
100	6	3	0	0	0
100	7	6	0	0	2
100	8	6	0	0	2
100	9	11	1	1	3
100	10	24	3	3	10
100	11	0	0	0	0
100	12	4	0	0	0
100	13	3	0	0	0
100	14	4	0	0	0
100	15	99	99	99	99
100	16	2	0	0	0
100	17	3	0	0	1
100	18	7	1	1	1
100	19	7	0	0	2
100	20	85	30	33	54
100	21	6	0	0	1
100	22	7	1	1	2
100	23	10	3	3	4
100	24	14	1	1	5
100	25	98	84	84	93
100	26	8	1	1	3
100	27	9	1	1	3
100	28	13	2	2	6
100	29	5	1	1	4
100	30	9	0	0	1
100	31	7	0	0	0
100	32	3	0	0	0
100	33	9	2	2	3
100	34	12	1	1	1
100	35	99	89	90	95
100	36	9	0	0	4
100	37	17	1	2	8
100	38	12	1	2	5
100	39	14	1	1	6
100	40	16	2	2	6

Condition 59(40 items test length, 1000/1000, and group mean difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	3	0	0	2
100	2	1	0	0	0
100	3	6	0	0	1
100	4	2	0	0	0
100	5	97	76	77	92
100	6	7	0	0	3
100	7	9	1	1	5
100	8	4	0	0	1
100	9	11	2	3	6
100	10	20	2	2	7
100	11	2	0	0	0
100	12	4	0	0	1
100	13	6	0	0	1
100	14	4	0	0	0
100	15	99	99	99	99
100	16	4	1	1	1
100	17	10	0	0	3
100	18	5	0	0	0
100	19	9	0	1	2
100	20	81	19	20	43
100	21	8	2	2	4
100	22	15	1	1	6
100	23	10	0	0	6
100	24	12	1	1	5
100	25	99	64	64	86
100	26	12	2	2	4
100	27	13	3	3	7
100	28	19	5	5	10
100	29	3	2	2	2
100	30	6	1	1	2
100	31	7	0	0	1
100	32	4	0	0	2
100	33	18	2	3	7
100	34	11	0	0	4
100	35	97	89	90	94
100	36	15	1	1	6
100	37	17	2	2	7
100	38	18	3	3	8
100	39	19	1	1	8
100	40	14	3	3	5

Condition 60(40 items test length, 1000/1000, and group SD difference)

Replication	item	Before	Bonferroni	Holms	BH
100	1	7	0	0	1
100	2	11	1	1	4
100	3	6	2	2	2
100	4	6	1	1	2
100	5	17	4	4	10
100	6	9	4	5	6
100	7	6	0	0	1
100	8	10	5	6	8
100	9	8	1	1	3
100	10	6	1	1	3
100	11	0	0	0	0
100	12	1	0	0	1
100	13	11	3	3	6
100	14	6	4	4	4
100	15	98	98	98	98
100	16	9	4	4	4
100	17	13	4	4	6
100	18	16	4	6	9
100	19	13	7	8	9
100	20	97	92	93	96
100	21	11	3	3	7
100	22	11	5	5	6
100	23	10	1	1	5
100	24	9	4	4	6
100	25	60	10	10	24
100	26	13	5	5	9
100	27	12	2	2	4
100	28	12	5	5	8
100	29	7	0	0	0
100	30	8	1	1	2
100	31	15	4	4	9
100	32	17	4	4	12
100	33	21	9	9	12
100	34	19	2	4	6
100	35	96	84	85	90
100	36	31	12	12	19
100	37	9	3	3	6
100	38	21	2	2	6
100	39	25	11	12	17
100	40	76	46	47	59

APPENDIX H

SAS CODE OF DATA GENERATION

```
options linesize=72;
proc datasets lib=work kill nolist memtype=data;
quit;
DM 'CLEAR OUTPUT';


%macro gen(sz1=, sz2=, length=, foc=, ref=, condition=);

      libname gen40me "D:\Simulation\gen 40 items no diff\";

      %do i=1 %to 100; *number of replications;

/* Focal Data Generation*/
      data focc;
            array item item1-item&length;
            array a a1-a&length;  /*tlength= test length*/
            array b b1-b&length;
            array p p1-p&length;
            array x x1-x&length;
            array y y1-y&length;
            infile 'D:\My Dropbox\sas\Focal40.dat'; /*parameter a and b
reference*/
         * infile 'D:\My Dropbox\sas\Focal20.dat';  /*parameter a and b
Focal (DIF Magnitude=10%)*/

                  do over a;
                     input item a b;
                  end;

                  do k=1 to &sz1;   /* set sample size*/
                     c=.20;
                     t1=rannor(0);  /*No group mean difference & No
group SD difference*/

               foc=1;

            do over p;
                     x=ranuni(0);
                     p=c+((1-c)/(1+exp(-1.7*a*(t&condition-b))));
                     if x le p then y = 1; else y=0;
                  end; output;
                  end;
       run;

      data gen40me.foc&foc&i; set focc; run;
      data ref;
            array item item1-item&length;
```

```
            array a a1-a&length;  /*tlength= test length*/
            array b b1-b&length;
            array p p1-p&length;
            array x x1-x&length;
            array y y1-y&length;
       infile 'D:\My Dropbox\sas\Reference40.dat'; /*parameter a and b
reference*/
      *     infile  'D:\My Dropbox\sas\Reference20.dat'; /*parameter a
and b reference*/

                  do over a;
                        input item a b;
                  end;

                  do k=1 to &sz1;   /* set sample size*/
                           c=.20;
                     foc=0;
                     t=rannor(0);

                  do over p;
                     x=ranuni(0);
                     p=c+((1-c)/(1+exp(-1.7*a*(t-b))));
                     if x le p then y = 1; else y=0;
                  end; output;
                  end;
       run;

       data gen40me.ref&ref&i; set ref; run;


proc printto; run;
%end;
%mend gen;
%gen (sz1=500, sz2=500, length=40,foc=1, ref=1, condition=1 );
%gen (sz1=1000, sz2=1000, length=40,foc=2, ref=2, condition=1 );
%gen (sz1=300, sz2=700, length=40,foc=3, ref=3, condition=1 );
%gen (sz1=500, sz2=1500, length=40,foc=4, ref=4, condition=1 );
```

SAS CODE OF NON-PARAMETRIC METHODS

```
options linesize=72;
libname MH "D:\Simulation\Mantel Haenszel\MH40\no\";
libname logist "D:\simulation\logistic Regression\";
libname gen40 "D:\Simulation\generation\items 40\gen 40 \";
libname gen20 "D:\simulation\gen20\";

proc datasets lib=work kill nolist memtype=data;
quit;
DM 'CLEAR OUTPUT';

%macro dissertation(sz1=, sz2=, length=, foc=, ref=, case=);
%do i=1 %to 5;
      data theta&i; set gen40.foc&foc&i gen40.ref&ref&i;

data intervals&i; set theta&i;

 /*Monte Carlo generation of data*/
       totscore=sum(of y1-y&length);
if totscore >=((20*&length)/100)-1 AND totscore =<((20*&length)/100)
THEN equint = 1;
else if totscore >=((20*&length)/100)+1 AND totscore
=<((30*&length)/100) THEN equint = 2;
else if totscore >=((30*&length)/100)+1 AND totscore
=<((40*&length)/100) THEN equint = 3;
else if totscore >=((40*&length)/100)+1 AND totscore
=<((50*&length)/100) THEN equint = 4;
else if totscore >=((50*&length)/100)+1 AND totscore
=<((60*&length)/100) THEN equint = 5;
else if totscore >=((60*&length)/100)+1 AND totscore
=<((70*&length)/100) THEN equint = 6;
else if totscore >=((70*&length)/100)+1 AND totscore
=<((80*&length)/100) THEN equint = 7;
run;

proc sort data=intervals&i;
by equint; run;

filename junk dummy;
proc printto  log=junk; run;
/***************************/
/****Mantel Haenszel Method****/
/***************************/

%do j=1 %to &length; /* number of items proc frequency*/
      *ods trace on/ label listing;
```

```
      ods trace off;
      ods exclude CommonRelRisks;
      ods exclude CrossTabFreqs;
      ods exclude BreslowDayTest;
      ods exclude Freq.Table1.CMH;
      ods output Freq.Table1.CMH=CMH&j;
      *ods listing; *shows output in output window;
      Proc Freq data=intervals&i;
                  tables equint*foc*y&j/CMH; *******GMH***********;
run;

      data Mantel&j; set CMH&j;
           if statistic=2 then output;
       run;
      proc append base=MHantel&i data=Mantel&j;  run;
      proc sort data=MHantel&i; by prob; run;
%end;
/*********************************/
/****Logistic Regression Procedure****/
/*********************************/
ods exclude ResponseProfile;
ods exclude ConvergenceStatus;
ods exclude FitStatistics;
ods exclude GlobalTests;
ods exclude OddsRatios;
ods exclude Association;
ods exclude ParameterEstimates;
ods output logist.ParameterEstimates=reg;

proc logist data=intervals&i;
     model y&j=totscore foc;
run;
ods listing;


data logist&j; set reg;
     rep=&i;
       item=&j;
       if Variable='foc' then output;
run;

proc append base=LR&i data=logist&j;* (where=(rep=&i));

proc sort data=LR&i; by decending Probchisq ; run;
%end;
/*****************************************************************
******/
```

```
data arrange&i; set MHantel&i;   /* set LR&I for logistic regression
procedure*/
      rep=&i;
       question=_n_;
             if Prob <=.05 then unadjusted=1; else unadjusted=0;
       if Prob <=(.05/&length) then Bonferroni=1; else Bonferroni=0;
       if Prob <=(.05/(&length-question+1)) then Holm=1; else Holm=0;
       if Prob <=((.05*question)/&length) then BH=1; else BH=0;

data summary&case; set arrange&case(keep=rep question unadjusted
Bonferroni Holm BH);
proc sort data= summary&case; by question rep; run;

data sum&case; set summary&case;
      by question rep;
      if first.question then unadj=0;
      unadj+unadjusted;
       if first.question then Bonf=0;
      Bonf+Bonferroni;
       if first.question then Holms=0;
      Holms+Holm;
      if first.question then B_H=0;
      B_H+BH;
      if last.question then output;
run;

data sum&case; set sum&case (drop=unadj Bonf Holms B_H); run;
proc transpose data=sum&case out=difsetting&case; run;

data difree&case; set difsetting&case (drop=col5 col10 col15
col20 );run;
proc transpose data=difree&case out=nondif&case; run;
data diff&case; set difsetting&case (keep=col5 col10 col15 col20 );run;
proc transpose data=diff&case out=dif&case; run;

proc printto; run;
%end;
%mend dissertation;

/*****************************************************************/
%dissertation (sz1=500, sz2=500, length=40,foc=1, ref=1, case=1);
```

SAS CODE OF PARAMETRIC METHOD

```
ibname Lord20 "D:\Simulation\Lords\Lord40\no\First stage\";
libname Second "D:\Simulation\Lords\Lord40\no\Second stage\";
*DM 'Clear log' ;
DM 'Clear Out' ;
%macro gen(dir=, filename=, length=);

%do j=1 %to 100;
data cov&j;
    infile "&dir.foc&filename&j..cov" firstobs=3 missover;
      input id 1-5
                  item $ 6-13
                  test $ 14-20
                  group 21
                  a
                  b
                  c
                  avar
                  abcov
                  /
                  bvar
                  accov
                  bccov
                  cvar;

                  asd=sqrt(avar);
                  bsd=sqrt(bvar);
                  csd=sqrt(cvar);
            c=.20;

data linking&j; set Lord20.Lordschi&j;

data linking&j; set linking&j(keep=id sig);

data foccov&j; set cov&j; set linking&j;
by id;


data foccov&j; set foccov&j;
if sig='1' then delete;

data focco&j; set foccov&j(keep=id item test group a b c avar abcov
bvar accov bccov cvar asd bsd csd);


data covt&j; set cov&j (rename=(avar=avarf bvar=bvarf abcov=covabf));
data covt&j; set covt&j(keep=avarf bvarf covabf);
stunbr=0000;


data sco&j;
/*In parentheses below, user must enter the name of their focal group
file with the .sco extension*/
```

```
      infile "&dir.foc&filename&j..sco" missover firstobs=3;  /*select
sample*/
      input group
                  id
                  /
                  resp 1-6
                  calib 7-7
                  subtest $ 8-15
                  attempt 16-20
                  correct 21-25
                  percent 26-35
                  theta 36-47
                  stderr 48-59
                  stdunest 60-60
                  grpprob 61-70
                  margprob 71-80;


data covref&j;
/*In parentheses below, user must enter the name of their reference
group file with the .cov extension*/
      infile "&dir.ref&filename&j..cov" missover firstobs=3;
      input id 1-5
                  item $ 6-13
                  test $ 14-20
                  group 21
                  a
                  b
                  c
                  avar
                  abcov
                  /
                  bvar
                  accov
                  bccov
                  cvar;

                  asd=sqrt(avar);
                  bsd=sqrt(bvar);
                  csd=sqrt(cvar);
                  c=.20;

data linking&j; set Lord20.Lordschi&j;

data linking&j; set linking&j(keep=id sig);

data refcov&j; set covref&j; set linking&j;
by id;

data refcov&j; set refcov&j;
if sig='1' then delete;

data refcovv&j; set refcov&j(keep=id item test group a b c avar abcov
bvar accov bccov cvar asd bsd csd);
```

```
data covreft&j; set covref&j (rename=(avar=avarr bvar=bvarr
abcov=covabr));
data covreft&j; set covreft&j (keep=avarr bvarr covabr);
stunbr=0000;


/* Built in mean and sigma linking.  For this research, Item 1 not
included in linking */
data lbase&j;
set  focco&j (firstobs=1);
proc means noprint mean std;
var b;
output out=outbase&j mean=mbase&j std=sbase&j;


data lg&j;
set refcovv&j (firstobs=2);
proc means noprint mean std;
var b;
output out=outg&j mean=mg&j std=sg&j;


data meansig&j;
merge outbase&j outg&j; by _type_;
alpha&j=sbase&j/sg&j;
beta&j=mbase&j-alpha&j*mg&j;
keep alpha&j beta&j;

proc printto; run;

/*Creating data sets to call into IML*/

data foc&j (keep = a b c avar abcov bvar accov bccov cvar); set cov&j;
data theta&j (keep = theta); set sco&j;
data ref&j (keep =  a b c avar abcov bvar accov bccov cvar); set
covref&j;
data link&j (keep = alpha&j beta&j) ; set meansig&j;

/*******************************************/
/*******************************************/
/*****************Lord's Chi-Square Test******/
/* proc iml;
**Creates a matrix with original focus group item parameter
information**;
use foc&j;
read all into matfoc&j;
*print matfoc&j;
**Creates a matrix with original focus group theta values and standard
error**;
use theta&j;
```

```
read all into mattheta&j;
**Creates a matrix with original reference group item parameter
information**;
use ref&j;
read all into matref&j;
**Creates a matrix with alpha linking coefficients**;
use link&j;
read all into matlink&j;
**Creates a matrix with beta linking coefficients**;
*/
**Values/Matrices to be used later**;
a_focc&j=repeat(0,nrow(matfoc&j),1);
b_foc&j=repeat(0,nrow(matfoc&j),2);
a_ref&j=repeat(matref&j,1,1);
b_ref&j=repeat(matref&j,1,2);
a_diff&j=repeat(0,&length,1);
b_diff&j=repeat(0,&length,1);
c_diff&j=repeat(0,&length,1);
origfoc&j=repeat(matfoc&j,1,1);
origref&j=repeat(matref&j,1,1);
diff&j=repeat(0,&length,3);
link&j=repeat(matlink&j,&length,1);

      a_focc&j=(1/link&j[1,1])*origref&j[,1];
      b_foc&j=link&j[1,1]*origref&j[,2]+link&j[1,2];
    a_diff&j=a_focc&j[,1]-origfoc&j[,1];
    b_diff&j=b_foc&j[,1]-origfoc&j[,2];
    c_diff&j=origfoc&j[,3]-origref&j[,3];

    diff&j=a_diff&j||b_diff&j||c_diff&j;
    dif&j=a_diff&j||b_diff&j;



create dif&j from dif&j;
append from dif&j;

quit;

data dif&j; set dif&j(rename=(col1=acom col2=bcom));
stunbr=0000;

data DIFFVECTOR&j; set covreft&j; set covt&j;
by stunbr;

data DIFFVECTOR&j; set DIFFVECTOR&j; set dif&j;
      by stunbr;
      varcom=avarr+avarf;
      covabcom=covabr+covabf;
      cvabcom2=(covabcom*covabcom);
      vrbcom=bvarr+bvarf;

Data Second.lordschi&j;
  set DIFFVECTOR&j;
```

```
        rep=&j;
            rank=_n_;
            DET=(varcom*vrbcom) - cvabcom2;
            LChi2=((acom*((acom*vrbcom)-(bcom*covabcom))) +
(bcom*((bcom*varcom)-(acom*covabcom))))/DET;
            prob=1-probchi(LChi2,2);
            if prob<0.05 then sig=1; else sig=0;
            *Keep rep rank prob sig;

proc sort data=Second.lordschi&j out=Second.arrange&j;
by prob;

data Second.arrange&j; set Second.arrange&j(keep=rep rank prob sig);
      id=_n_;
    if prob <= 0.05 then unadjusted=1; else unadjusted=0;
    if prob <= 0.05/40 then Bonferroni=1; else Bonferroni=0;
      if prob <= (0.05/(40-id+1)) then Holm=1; else Holm=0;
      if (prob*(40/id)) <= 0.05 then BH=1; else BH=0;

data Second.arrange&j; set Second.arrange&j(keep=id rep rank prob
unadjusted Bonferroni Holm BH);


proc datasets lib=work nolist;
delete all&j;

%END;
proc printto; run;
%mend gen;
/*************************************/
%gen (dir=D:\Simulation\Bilog40\no\, filename=2, length=40);
/*************************************/
/*************************************/
/****************DFIT*****************/
/*************************************/
proc iml;
**Creates a matrix with original focus group item parameter
information**;
use orig;
read all into matorig;
**Creates a matrix with original focus group theta values and standard
error**;
use theta;
read all into mattheta;
**Creates a matrix with original reference group item parameter
information**;
use ref;
read all into matref;
**Creates a matrix with alpha and beta linking coefficients**;
use link;
read all into matlink;

**Values/Matrices to be used later**;
seeds={123456 234567 345678 456789 567890 678901};
```

```
items=nrow(matorig);
n=nrow(mattheta);
reps=1000;
ncdifmat=repeat(0,reps,items);
dtfmat=repeat(0,reps,1);
fnor=repeat(0,3,items);
rnor=repeat(0,3,items);
fnort=repeat(0,3,items);
rnort=repeat(0,3,items);
foc=repeat(0,3,items);
ref=repeat(0,3,items);
pfoc=repeat(0,n,items);
pref=repeat(0,n,items);
T=repeat(0,3,3);
r=repeat(1,3,3);


** 1 Parameter
Model*************************************************************;
if (matorig[:,9]=0 & matorig[:,7]=0) then do;
      do rep=1 to reps;
            do i=1 to items;
                  do param=1 to 3;
                        **Creates random normally distributed item
parameters for focal and reference groups**;
                        fnor[param,i]=normal(seeds[1,param]*i+rep);
                        rnor[param,i]=normal(seeds[1,3+param]*i+rep);
                  end;
            end;

            do i=1 to items;
                  do param=1 to 3;
                        **Changes normal matrices to have same means
and standard deviations as originals**;
                        **These will be the final simulated item
parameters used to calculate p**;

      foc[param,i]=matorig[i,param]+(matorig[i,6+param]*fnor[param,i]);

ref[param,i]=matorig[i,param]+(matorig[i,6+param]*rnor[param,i]);
                 end;
            end;

            do theta=1 to n;
                  do i=1 to items;
                        **Calculates p for each set of item parameters
using thetas from BILOG**;
                     pfoc[theta,i]=foc[3,i]+(1-foc[3,i])*
                             ((EXP(1.7*foc[1,i]*(mattheta[theta,1]-
foc[2,i])))/
                             (1+EXP(1.7*foc[1,i]*(mattheta[theta,1]-
foc[2,i]))));
                     pref[theta,i]=ref[3,i]+(1-ref[3,i])*
```

```
                              ((EXP(1.7*ref[1,i]*(mattheta[theta,1]-
ref[2,i])))/
                              (1+EXP(1.7*ref[1,i]*(mattheta[theta,1]-
ref[2,i]))));
                  end;
            end;

            **Calculates d used in NCDIF equation**;
            d=pfoc-pref;

            **Calculates sum of d (capital d) used in DTF equation**;
            sumd=d[,+];

            **Calculates NCDIF**;
            do i = 1 to items;
                  ncdifmat[rep,i]=((sum(d[##,i])-
(((d[+,i])**2)/(n)))/(n))+((d[:,i])**2);
            end;

            **Calculates DTF**;
            dtfmat[rep,1]=((sum(sumd[##,1])-
(((sumd[+,1])**2)/(n)))/(n))+((sumd[:,1])**2);
      end;
end;


*************************************************************************
*****************
**Two Parameter Model and Three Parameter Model with a Fixed c**;
else if (matorig[:,9]=0 & matorig[:,7]<>0) then do;
      do rep=1 to reps;
            do i=1 to items;
                  **Fills r then makes T if the r matrix is positive
definite**;
                  r[1,2]=matorig[i,4]/(matorig[i,7]*matorig[i,8]);
                  r[2,1]=matorig[i,4]/(matorig[i,7]*matorig[i,8]);
                  r[1,3]=0;
                  r[3,1]=0;
                  r[2,3]=0;
                  r[3,2]=0;
                  T=half(r);

                  do param=1 to 3;
                        **Creates random normally distributed item
parameters for focal and reference groups**;
                        fnor[param,i]=normal(seeds[1,param]*i+rep);
                        rnor[param,i]=normal(seeds[1,3+param]*i+rep);
                  end;

                  **Transforms simulated item parameters to have same
covariances as originals**;
                  fnort[,i]=T`*fnor[,i];
                  rnort[,i]=T`*rnor[,i];
            end;
```

```
            do i=1 to items;
                  do param=1 to 3;
                        **Changes normal matrices to have same means
and standard deviations as originals**;
                        **These will be the final simulated item
parameters used to calculate p**;

foc[param,i]=matorig[i,param]+(matorig[i,6+param]*fnort[param,i]);

ref[param,i]=matorig[i,param]+(matorig[i,6+param]*rnort[param,i]);
                 end;
          end;

            do theta=1 to n;
                  do i=1 to items;
                        **Calculates p for each set of item parameters
using thetas from BILOG**;
                      pfoc[theta,i]=foc[3,i]+(1-foc[3,i])*
                              ((EXP(1.7*foc[1,i]*(mattheta[theta,1]-
foc[2,i])))/
                              (1+EXP(1.7*foc[1,i]*(mattheta[theta,1]-
foc[2,i]))));
                      pref[theta,i]=ref[3,i]+(1-ref[3,i])*
                              ((EXP(1.7*ref[1,i]*(mattheta[theta,1]-
ref[2,i])))/
                              (1+EXP(1.7*ref[1,i]*(mattheta[theta,1]-
ref[2,i]))));
                   end;
            end;

            **Calculates d used in NCDIF equation**;
            d=pfoc-pref;

            **Calculates sum of d (capital d) used in DTF equation**;
            sumd=d[,+];

            **Calculates NCDIF**;
            do i = 1 to items;
                  ncdifmat[rep,i]=((sum(d[##,i])-
(((d[+,i])**2)/(n)))/(n))+((d[:,i])**2);
            end;

            **Calculates DTF**;
            dtfmat[rep,1]=((sum(sumd[##,1])-
(((sumd[+,1])**2)/(n)))/(n))+((sumd[:,1])**2);
      end;
end;


************************************************************************
************;
************************************************************************
************;
```

```
**Three Parameter Model without Fixed c**;
else if (matorig[:,9]<>0 & matorig[:,7]<>0) then do;
      problem_c=repeat('         ',1,items);

      do rep=1 to reps;
            do i=1 to items;
                  **Fills r then makes T if the r matrix is positive
definite**;
                  r[1,2]=matorig[i,4]/(matorig[i,7]*matorig[i,8]);
                  r[2,1]=matorig[i,4]/(matorig[i,7]*matorig[i,8]);
                  r[1,3]=matorig[i,5]/(matorig[i,7]*matorig[i,9]);
                  r[3,1]=matorig[i,5]/(matorig[i,7]*matorig[i,9]);
                  r[2,3]=matorig[i,6]/(matorig[i,8]*matorig[i,9]);
                  r[3,2]=matorig[i,6]/(matorig[i,8]*matorig[i,9]);

                  if det(r)>0 then do;
                        T=half(r);
                  end;

                  if det(r)<=0 then do;
                        problem_c[1,i]='x        ';
                        r[1,2]=matorig[i,4]/(matorig[i,7]*matorig[i,8]);
                        r[2,1]=matorig[i,4]/(matorig[i,7]*matorig[i,8]);
                        r[1,3]=0;
                        r[3,1]=0;
                        r[2,3]=0;
                        r[3,2]=0;
                        T=half(r);
                  end;

                  do param=1 to 3;
                        **Creates random normally distributed item
parameters for focal and reference groups**;
                        fnor[param,i]=normal(seeds[1,param]*i+rep);
                        rnor[param,i]=normal(seeds[1,3+param]*i+rep);
                  end;

                  **Transforms simulated item parameters to have same
covariances as originals**;
                  fnort[,i]=T`*fnor[,i];
                  rnort[,i]=T`*rnor[,i];
            end;

            do i=1 to items;
                  do param=1 to 3;
                        **Changes normal matrices to have same means
and standard deviations as originals**;
                        **These will be the final simulated item
parameters used to calculate p**;

foc[param,i]=matorig[i,param]+(matorig[i,6+param]*fnort[param,i]);

ref[param,i]=matorig[i,param]+(matorig[i,6+param]*rnort[param,i]);
                end;
```

```
          end;

            do theta=1 to n;
                  do i=1 to items;
                        **Calculates p for each set of item parameters
using thetas from BILOG**;
                     pfoc[theta,i]=foc[3,i]+(1-foc[3,i])*
                              ((EXP(1.7*foc[1,i]*(mattheta[theta,1]-
foc[2,i])))/
                              (1+EXP(1.7*foc[1,i]*(mattheta[theta,1]-
foc[2,i]))));
                     pref[theta,i]=ref[3,i]+(1-ref[3,i])*
                              ((EXP(1.7*ref[1,i]*(mattheta[theta,1]-
ref[2,i])))/
                              (1+EXP(1.7*ref[1,i]*(mattheta[theta,1]-
ref[2,i]))));
                  end;
            end;

            **Calculates d used in NCDIF equation**;
            d=pfoc-pref;

            **Calculates sum of d (capital d) used in DTF equation**;
            sumd=d[,+];

            **Calculates NCDIF**;
            do i = 1 to items;
                  ncdifmat[rep,i]=((sum(d[##,i])-
(((d[+,i])**2)/(n)))/(n))+((d[:,i])**2);
            end;

            **Calculates DTF**;
            dtfmat[rep,1]=((sum(sumd[##,1])-
(((sumd[+,1])**2)/(n)))/(n))+((sumd[:,1])**2);
      end;

      title3 ' ';
      print 'Columns marked with x are items with simulated c-
parameters not related to a and b' problem_c;

end;

************************************************************************
***********;
************************************************************************
***********;

**Creates an itemrank matrix with ncdif values for each item in
ascending order**;
itemrank=repeat(0,reps,items);
do i=1 to items;
      k=repeat(0,reps,1);
      k=ncdifmat[,i];
      f=k;
```

```
      k[rank(k),]=f;
      itemrank[,i]=k;
end;
create difcut40.foc214 from ncdifmat;
append from ncdifmat;

**Creates a testrank matrix with dtf values in ascending order**;
testrank=dtfmat;
ff=dtfmat;
dtfmat[rank(dtfmat),]=ff;
testrank[,1]=dtfmat;

title3 ' ';
cutoffnames={'Cutoff .10', 'Cutoff .05', 'Cutoff .01', 'Cutoff .001'};
NCDIF_ITEM_CUTOFFS=repeat(0,4,items);
NCDIF_ITEM_CUTOFFS[1,]=itemrank[ceil(.90*reps),];
NCDIF_ITEM_CUTOFFS[2,]=itemrank[ceil(.95*reps),];
NCDIF_ITEM_CUTOFFS[3,]=itemrank[ceil(.99*reps),];
NCDIF_ITEM_CUTOFFS[4,]=itemrank[ceil(.999*reps),];
*print NCDIF_ITEM_CUTOFFS [r=cutoffnames];

**Creates an empty column matrix that will be filled with NCDIF
values**;
ncdifcol=repeat(0,reps*items,1);

**Reads NCDIF values 1 column**;
do i=1 to items;
      do r=1 to reps;
            ncdifcol[r+(i-1)*reps,1]=ncdifmat[r,i];
      end;
end;

**Puts the NCDIF values in rank order**;
b=ncdifcol;
ncdifcol[rank(ncdifcol),]=b;


**Computes cutoff scores at the .001, .01, .05, and .10 levels**;
x=nrow(ncdifcol);
NCDIF_TOTAL_CUTOFFS=repeat(0,4,1);
NCDIF_TOTAL_CUTOFFS[1,1]=ncdifcol[ceil(.90*x),1];
NCDIF_TOTAL_CUTOFFS[2,1]=ncdifcol[ceil(.95*x),1];
NCDIF_TOTAL_CUTOFFS[3,1]=ncdifcol[ceil(.99*x),1];
NCDIF_TOTAL_CUTOFFS[4,1]=ncdifcol[ceil(.999*x),1];
*print NCDIF_TOTAL_CUTOFFS [r=cutoffnames];

**Prints cutoff scores for DTF;
DTF_CUTOFFS=repeat(0,4,1);
DTF_CUTOFFS[1,]=testrank[ceil(.90*reps),];
DTF_CUTOFFS[2,]=testrank[ceil(.95*reps),];
DTF_CUTOFFS[3,]=testrank[ceil(.99*reps),];
DTF_CUTOFFS[4,]=testrank[ceil(.999*reps),];
*print DTF_CUTOFFS [r=cutoffnames];
```

```
*************************************************************************
*************
**Puts the reference group on the same scale as the focal group**;
newref=repeat(0,items,3);
do i=1 to items;
      newref[i,1]=(1/matlink[1,1])*matref[i,1];
      newref[i,2]=matlink[1,1]*matref[i,2]+matlink[1,2];
      newref[i,3]=matref[i,3];
end;

**Calculates p for the focal group and linked reference group**;
pf=repeat(0,n,items);
pr=repeat(0,n,items);
NCDIF=repeat(0,1,items);
      do theta=1 to n;
            do i=1 to items;
                  **Calculates p for each set of item parameters using
thetas from BILOG**;
               pf[theta,i]=matorig[i,3]+(1-matorig[i,3])*
                        ((EXP(1.7*matorig[i,1]*(mattheta[theta,1]-
matorig[i,2])))/
                        (1+EXP(1.7*matorig[i,1]*(mattheta[theta,1]-
matorig[i,2]))));
               pr[theta,i]=newref[i,3]+(1-newref[i,3])*
                        ((EXP(1.7*newref[i,1]*(mattheta[theta,1]-
newref[i,2])))/
                        (1+EXP(1.7*newref[i,1]*(mattheta[theta,1]-
newref[i,2]))));
            end;
      end;

**Calculates d used in NCDIF equation**;
d=pf-pr;

**Calculates sum of d (capital d) used in DTF equation**;
sumd=d[,+];

**Calculates NCDIF**;
do i = 1 to items;
    NCDIF[1,i]=((sum(d[##,i])-(((d[+,i])**2)/(n)))/(n))+((d[:,i])**2);
end;
create difcut40.NCDIF214 from NCDIF;
append from NCDIF;

**Calculates DTF**;
DTF=repeat(0,1,1);
DTF[1,1]=((sum(sumd[##,1])-
(((sumd[+,1])**2)/(n)))/(n))+((sumd[:,1])**2);
print NCDIF;

**Flags significant NCDIF**;
sig_NCDIF=repeat('         ',1,items);
do i=1 to items;
```

```
      if NCDIF[1,i]>=NCDIF_ITEM_CUTOFFS[1,i] then sig_NCDIF[1,i]='*
';
      if NCDIF[1,i]>=NCDIF_ITEM_CUTOFFS[2,i] then sig_NCDIF[1,i]='**
';
      if NCDIF[1,i]>=NCDIF_ITEM_CUTOFFS[3,i] then sig_NCDIF[1,i]='***
';
      if NCDIF[1,i]>=NCDIF_ITEM_CUTOFFS[4,i] then sig_NCDIF[1,i]='****
';
      if NCDIF[1,i]<NCDIF_ITEM_CUTOFFS[1,i] then sig_NCDIF[1,i]='ns
';
end;

print sig_NCDIF;

*print DTF;

**Flags significant DTF**;
sig_DTF=repeat('         ',1,1);
      if DTF[1,1]>=DTF_CUTOFFS[1,1] then sig_DTF[1,1]='*        ';
      if DTF[1,1]>=DTF_CUTOFFS[2,1] then sig_DTF[1,1]='**       ';
      if DTF[1,1]>=DTF_CUTOFFS[3,1] then sig_DTF[1,1]='***      ';
      if DTF[1,1]>=DTF_CUTOFFS[4,1] then sig_DTF[1,1]='****     ';
      if DTF[1,1]<DTF_CUTOFFS[1,1] then sig_DTF[1,1]='ns       ';

*print sig_DTF;

quit;

run;

/*************************************************************************/
/*************************************************************************/
```