The Effects of Matching Criterion Contamination on the Mantel-Haenszel Procedure

by

Randall David Penfield

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Curriculum, Teaching and Learning
Ontario Institute for Studies in Education of the University of Toronto

© Randall David Penfield 2000
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.
The Effects of Matching Criterion Contamination on the Mantel-Haenszel Procedure

by

Randall David Penfield

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Curriculum, Teaching & Learning
University of Toronto
© Randall David Penfield 2000

ABSTRACT

Modern bias detection procedures search for differences in item performance between demographic groups after conditioning on an estimate of the ability intended to be measured by the test. The estimate of ability is typically some function of the total test score. Since examinees with equal test scores are said to be matched on ability, the internal measure of ability is referred to as the matching criterion. When the test contains one or more biased items, the test score will not be a valid measure of ability. As a result, the matching criterion is said to be contaminated by the biased items.

This study consists of a comprehensive examination of the Mantel-Haenszel (MH) procedure in the presence of a contaminated matching criterion. This examination focused on two primary issues: (1) assessing the effects of contamination on the MH procedure, and (2) developing alternative DIF detection procedures which are robust to contamination.

The results indicate that the presence of contamination has minor effects on the MH procedure when contamination is small or moderate, but has substantial effects when
contamination is large. The effects were related to the proportion of items containing contamination, and increased as the level of DIF in the contaminated items increased.

Two solutions to the problem of matching criterion contamination solutions were proposed. First, a procedure was developed that adjusts the obtained MH value to correct for the effects of contamination. The results of a simulation study suggest that the adjustment is effective in general at correcting for the effects of contamination, losing efficiency only under the most severe levels of contamination and the smaller sample size \((N = 250)\). A second solution to the problem of contamination was the proposal of \(MB-DIF\), a new statistic that is theoretically robust to the effects of contamination. The results of a simulation study indicate that the performance of \(MB-DIF\) exceeds that of the MH adjustment, particularly when sample sizes were large \((N = 1000)\). Under the condition of large sample sizes, \(MB-DIF\) was completely robust to the effects of contamination, maintaining power and Type I error rates identical to control conditions in which no contamination existed. The performance of \(MB-DIF\) suffered slightly when sample sizes were small \((N = 250)\), largely due to inflated Type I error rates under large levels of contamination.

The findings have two implications. First, bias detection analyses should consider the possible magnitude of bias in other items in the test when investigating the magnitude of bias in any given item. Second, adjustment procedures can control for the majority of the underestimation in DIF statistics when the matching criterion is contaminated. It appears that \(MB-DIF\) offers a more effective solution to the problem of matching criterion contamination than adjusting the MH value.
Acknowledgments

First, I would like acknowledge the time and energy spent by my thesis committee. A special thanks to Dr. Ross Traub for his conceptual and editorial advice, Dr. Phil Nagy for the thorough review of the manuscript and those nicely typed comments, Dr. Ruth Childs for valuable advice on making the dissertation better organized, Richard Wolfe for pointing me in the right direction on several instances concerning the mathematical and statistical rigor of the manuscript, and Dr. Hariharan Swaminathan for his insightful recommendations. I would also like to thank Dr. Shizuhiko Nishisato for his early involvement in the proposal of the dissertation.

I would also like to acknowledge the support I received from numerous individuals during my doctoral studies. I am indebted to Phil Nagy for supplying seemingly endless opportunities for funded project research that afforded me the time, resources, and experience necessary to complete my studies in such timely fashion. A huge thanks to Alex Lawson for her partnership over the years, making the late nights in the office bearable, giving me something to look forward to in the evenings (no merlot please!), and bringing a sense of community to our group. I hope it doesn’t end here. Many thanks to Adrienne McRuvie for providing the laughter in my life, and not letting me forget the big picture. Your lessons will be a part of my life always. Finally, to my parents Doug and Kath for their influence and support throughout this chapter of my life. Individually you both supplied me with the motivation and insight required to get me where I am today. Thanks!
# Table of Contents

Abstract  ii  
Acknowledgments  iv  

Chapter One: Introduction  
General Overview of the Problem  1  
Defining Item Bias  1  
Measuring Item Bias Using Differential Item Functioning  3  
Methods for Measuring DIF  4  
Definition of the Problem: Contamination-Induced Misclassification  14  
Practical Implications of CIM on DIF Detection  19  
Previous Studies of the Effects of CIM  20  
Organization and Objectives  21  

Chapter Two: The Effects of CIM on the MH Procedure  
Introduction  23  
Method  24  
Results  27  
Discussion  34  

Chapter Three: Developing an Adjustment to the MH Procedure to Correct for CIM  
Introduction  39  
A Bivariate Latent Trait Model of Bias  39  
Relevant Assumptions and Results  42  
Developing the Adjustment  44  

Chapter Four: Estimating the Parameters of the Adjustment  
Introduction  49  
The Estimation Procedures  49  

Chapter Five: Testing the Performance of the MH Adjustment  
Introduction  59  
Method  59  
Results  62  
Discussion  71
Chapter Six: Developing and Assessing MB-DIF

Introduction
Defining MB-DIF

Chapter Seven: Assessing the Performance of MB-DIF

Introduction
Method
Results
Discussion

Chapter Eight: Practical Implications of Findings

Summary of the Major Findings
Practical Implications
Limitations to Practical Application

References

Appendices

A Simulation program for Study One
B Simulation program for Study Two
C Simulation program for Study Three
D Proof of the relationship between MB-DIF and STNDP-DIF
E Estimation procedure for the number of focal group members at each valid score category

Tables

1.1 Data for the jth Matched Set of R and F
1.2 Proportions for the jth Matched Set of R and F
2.1 Mean Values of MHD-DIF for t = 0.6, 20 Items
2.2 Mean Values of MHD-DIF for t = 0.6, 60 Items
2.3 Type I Error Rates of MH^2 for 20 Item Tests
2.4 Type I Error Rates of MH^2 for 60 Item Tests
4.1 Steps Used in Estimating Valid Focal Group Item Difficulty
5.1 Mean Adjusted and Unadjusted MHD-DIF Values, z-scores, Power and Proportion in Each ETS Category for Sample Sizes of 1000 Having Equal Mean Abilities
5.2 Mean Adjusted and Unadjusted \textit{MHD-DIF} Values, z-scores, Power and Proportion in Each ETS Category for Sample Sizes of 250 Having Equal Mean Abilities

5.3 Mean Adjusted and Unadjusted \textit{MHD-DIF} Values, z-scores, Power and Proportion in Each ETS Category for Sample Sizes of 1000 Having Unequal Mean Abilities

5.4 Mean Adjusted and Unadjusted \textit{MHD-DIF} Values, z-scores, Power and Proportion in Each ETS Category for Sample Sizes of 250 Having Unequal Mean Abilities

5.5 Type I Error Rates for Adjusted \textit{MH}\textsuperscript{2} for Group Size of 1000

5.6 Type I Error Rates for Adjusted \textit{MH}\textsuperscript{2} for Group Size of 250

7.1 Mean \textit{MB-DIF} Values, Type I Error Rates, and Power for Sample Sizes of 1000 Having Equal Mean Abilities

7.2 Mean \textit{MB-DIF} Values, Type I Error Rates, and Power for Sample Sizes of 250 Having Equal Mean Abilities

7.3 Mean \textit{MB-DIF} Values, Type I Error Rates, and Power for Sample Sizes of 1000 Having Unequal Mean Abilities

7.4 Mean \textit{MB-DIF} Values, Type I Error Rates, and Power for Sample Sizes of 250 Having Unequal Mean Abilities

Figures

2.1 The Effects of Contamination on DIF Detection

5.1 A Comparison of the Control, Observed and Adjusted Results
Chapter One

Introduction and Statement of the Problem

General Overview of the Problem

Modern bias detection techniques search for differences in item performance between demographic groups after conditioning on an estimate of the ability intended to be measured by the test. The estimate of ability is typically some function of the total test score. When the test contains one or more biased items, the test score will not be a valid measure of ability, resulting in an underestimation of the magnitude of bias in an item. The problem at hand is one of determining the extent to which the detection of bias in a given item is affected by the presence of other biased items in the test. The aim of this thesis is to investigate the full scope of this problem and offer a range of solutions.

This chapter is dedicated to defining the problem to be addressed by this dissertation. To this end, I will first define bias and summarize the modern techniques used to detect item bias. Next, I will present a mathematical proof that the presence of biased items in the estimate of ability causes bias detection statistics to underestimate the magnitude of bias in an item. This proof will be followed by a discussion of the practical implications of the problem. I will conclude this chapter with an outline of the work presented in the remainder of the dissertation.

Defining Item Bias

It is sometimes the case that a test item is phrased in such a way that it does not
accurately reflect the ability level of the concepts intended to be measured by the test for a certain population of examinees. Such an item requires sources of knowledge that are different from those intended to be measured by the test, causing it to be less valid for a particular group. As a result, such an item is said to be biased. Camilli and Shepard (1994) offer an effective analogy:

*Test bias* is defined as invalidity or systematic error in how a test measures for members of a particular group. Bias is systematic in the sense that it creates a distortion in test results for members of a particular group. This is analogous to clocking individuals to measure their running speed, but using a stopwatch that runs too slowly for black runners. In this example, the rankings within groups of black runners and white runners might be relatively accurate, but comparisons between the best runners from each group and comparisons between group averages would be confounded by a bias in the stopwatch (italics inset) (p. 8).

While this definition of bias has intuitive appeal, it lacks a rigorous conceptualization of the conditions necessary for bias to exist. Modern investigations into bias have operationalized bias as the impact of two simultaneous factors on item performance (Camilli, 1992; Kok, 1988; Shealy & Stout, 1993a, b). The first condition is that the correct response to a given item (referred to as the *studied item*) requires not only a certain level of ability in the content domain intended to be measured by the test, but also abilities or resources that are different from those intended to be measured. Using the terminology of Shealy and Stout (1993b), the abilities intended to be measured are called the *target abilities*, and those which are present but unintended are called the *nuisance determinants*. The second condition is that the levels of the nuisance determinants are different across two subgroups of the examinees being tested. The two subgroups are typically referred to as the *reference group* and the *focal group*; the reference group usually consists of the majority of examinees who perform the
studied item relatively better, and the focal group usually consists of the minority of examinees who perform the studied item relatively worse. The result of these two conditions is a situation whereby nuisance determinants are present in the test, and one group of examinees, typically the reference group, has higher levels of the nuisance determinants than the focal group, and so does relatively better than the focal group on the studied item. Stated differently, bias exists when a difference in item performance between the focal and reference groups cannot be attributed solely to differences in the target abilities, and thus must be accounted for, at least partially, by the nuisance determinants. For example, if a math question is stated in terms of baseball batting percentages, the males (reference group), who we expect to have a better knowledge of baseball jargon (nuisance determinant) than females (focal group), might also be expected to perform better on the item than females, even though both males and females may have equal math ability (target ability).

Measuring Item Bias Using Differential Item Functioning

Early bias detection techniques were based on assessing the differences in mean performances of demographic groups (Cleary & Hilton, 1968; Angoff, 1972, 1982). Unfortunately, it is impossible to tell from mean differences alone whether a difference in item performance is due to a difference in target ability, or a difference in the nuisance determinant. To disentangle the effects of target ability and nuisance determinant, modern bias detection techniques search for a difference in item performance for the reference and focal groups after conditioning on an estimate of the target ability (see Dorans & Holland, 1993). Total test score is commonly used as the estimate of target ability. Examinees in the
focal and reference groups who are estimated to possess the same level of target ability are said to be *matched* on that ability (e.g., focal and reference group examinees with a total test score of 20 might be said to be matched, as are those who obtained a score of 21). Assuming that the estimates of target ability are valid, any difference in performance that exists after matching cannot be attributed to a difference in target ability, and thus is explained by the alleged nuisance determinant. In this situation the item functions differently for the two groups, and as a result *differential item functioning* (DIF) is said to exist.

It is important to preserve a clear distinction between DIF and bias. While it is true that modern methods have addressed bias in the framework of DIF, it is logically incorrect to claim that the presence of DIF necessarily indicates the existence of bias. That is, the presence of DIF is a necessary, but not sufficient, condition for the presence of bias. The existence of bias requires the additional consideration of construct validity. Construct validity evidence must show that there is some other definable trait, other than that intended to be measured, which is causing the DIF (Camilli & Shepard, 1994, pp. 2, 16).

**Methods for Measuring DIF**

Many statistics have been developed for identifying DIF over the past two decades. It is generally agreed that DIF can be conceptualized best in the context of item response theory (IRT). IRT regresses the probability of correct response \( P \) on an estimate of the latent ability intended to be measured \( \theta \). The three parameter logistic regression IRT model (3PL) can be expressed as
where \(a\) is the discrimination parameter, \(b\) is the difficulty parameter, and \(c\) is the pseudo guessing parameter. The line represented by Equation 1.1 is commonly referred to as the item characteristic curve (ICC). More detailed accounts of the theory and application of item response models are given by Hambleton, Swaminathan and Rogers (1991), and Lord (1980).

The most common conceptualization of bias within an IRT framework is a simple difference in the difficulty parameter between groups, making the probability of success on the item lower for one group at any given level of ability. However, DIF need not be restricted to a difference in difficulty parameter, as a substantial difference in any one of the three parameters will cause the item to function differently for the two groups.

Given the clear conceptualization of DIF in an IRT framework, it is not surprising that a myriad of IRT-based DIF detection methods have been proposed. Lord (1977, 1980) was the first to suggest that differences in the IRT parameters between groups serve as an ideal indicator of DIF. He proposed two tests for evaluating the significance of DIF. The simpler test compares the difficulty parameters for the two groups. The significance of this difference can be tested by

\[
P = c + \frac{1 - c}{1 + e^{-1.7a(\theta - b)}},
\]

(1.1)

where the numerator is the difference between the estimated \(b\) parameters for the reference and focal groups, and the denominator is the standard error of the difference between these parameters. Since \(d\) is distributed approximately as a unit normal variable, a test of
significance is available. Lord (1980) noted that similar tests could be performed to investigate differences between the discrimination parameters as well.

An extension of this method is a chi-square test which tests for the joint difference between the difficulty and discrimination parameters for the two groups (Lord, 1980). The differences between estimated $a$ and $b$ parameters for the item in question can be represented by the vector

$$\hat{V}' = (\hat{a}_F - \hat{a}_R, \hat{b}_F - \hat{b}_R),$$

where $F$ and $R$ refer to the focal and reference groups. The test statistic can be expressed as

$$\chi^2 = \hat{V}'S^{-1}\hat{V}$$

(1.4)

where $S$ is the estimate of the sampling variance-covariance matrix of the differences between the item parameter estimates. Details concerning the computation of $S$ are given in Lord (1980, p. 223). The $c$ values are estimated using both the reference and focal group members together. The test statistic has an asymptotic chi-square distribution with two degrees of freedom. While this method has theoretical appeal, it has been shown to be ineffective in comparative research studies (Raju, Drasgow & Slinde, 1991, as cited in Camilli & Shepard, 1994). This ineffectiveness can be attributed primarily to its lack of consideration of the density of examinees along the ability continuum.

Since Lord’s work nearly two decades ago, substantial advances have been made in model fitting and hypothesis testing in the context of IRT. These advances have led to more sophisticated IRT-based DIF detection techniques, the most popular of which is the IRT likelihood ratio approach (Thissen, Stienberg & Wainer, 1988). This approach compares the
fit of an item response model having common parameters for the focal and reference groups (compact model) to one having separate parameters for each group (augmented model). The augmented model includes all of the parameters of the compact model, as well as additional parameters, such that the compact model is hierarchically nested within the augmented model. The log of the ratio of likelihoods can be used to test whether the additional parameters in the augmented model significantly improve the fit of the model. The form of the test is

\[ G^2(d.f.) = -2 \log \left[ \frac{\text{Likelihood(Augmented)}}{\text{Likelihood(Compact)}} \right] \] (1.5)

where \(d.f.\) is the difference between the number of parameters in the augmented and compact models. Under very general assumptions, the value of \(G^2(d.f.)\) is distributed as \(\chi^2(d.f.)\) under the null hypothesis of no difference between the models (Rao, 1973, pp. 418-420). The augmented model has separate \(b\)s for the reference and focal groups, while the compact model has a common \(b\), estimated from the combined reference and focal group samples.

Although IRT methods have theoretical appeal, they are constrained by sample size requirements, assumptions concerning model fit, and the software necessary to calibrate the items. An alternative to IRT methods is the logistic regression procedure to detect DIF, first proposed by Swaminathan and Rogers (1990). The logistic regression equation can be written as

\[
P(U|G) = \frac{e^{(\beta_0 + \beta m + \beta G + \beta m^* G)}}{1 + e^{(\beta_0 + \beta m + \beta G + \beta m^* G)}}
\] (1.6)

where \(U\) represents the dichotomous response to an item, \(\beta_0\) represents the weight associated
with the intercept, $\beta_i$ represents the weight attached to the matching variable $m$ (usually the total test score), $\beta_g$ represents the weight attached to the group variable $G$, and $\beta_s$ represents the weight attached to the group-by-score interaction $G*m$. Tests of significance of the coefficients $\beta_2$ and $\beta_3$ provide answers to the questions concerning uniform DIF (DIF favours one group over the entire ability range) and nonuniform DIF (DIF favours different groups at different regions of the ability range), respectively. The difference in the log of the likelihood functions obtained in regressions with and without the $\beta_3$ coefficient is used to test for nonuniform DIF. The difference in the log of the likelihood function obtained in regressions with and without the $\beta_2$ coefficient is used to test for uniform DIF. Compared to IRT methods of DIF detection, the logistic regression procedures offer greater ease because no estimate of the latent ability is required. Unfortunately, logistic regression methods still require maximum likelihood estimation of the model parameters, and as a consequence require substantial computer resources to perform the necessary computations.

The limitations of IRT and logistic regression procedures have led bias detection researchers to favour contingency table approaches that have the advantage of being computationally simple, due to the lack of any assumed functional relationships between the probability of correct response and ability level. These methods share the common characteristic that they use an observed test score as the matching variable, and then create a two-by-two contingency table, crossing group membership and item success, for each of the matched sets of examinees. To keep the notation used here consistent with that used in the remainder of the dissertation, let the total test score be denoted by $m$, and each level of the
total test score be denoted by $j$. The data for the $j$th matched set is displayed in Table 1.1, where $T_j$ is the total number of reference and focal group examinees in the $j$th matched set; $N_{Rj}$ is the number of these who are in the reference group, and of these $A_j$ answered the studied item correctly. The other entries in Table 1.1 have similar definitions.

<table>
<thead>
<tr>
<th>Score on Studied Item</th>
<th>1</th>
<th>0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>$A_j$</td>
<td>$B_j$</td>
<td>$N_{Rj}$</td>
</tr>
<tr>
<td>Focal</td>
<td>$C_j$</td>
<td>$D_j$</td>
<td>$N_{Fj}$</td>
</tr>
<tr>
<td>Total</td>
<td>$N_{ij}$</td>
<td>$N_{0j}$</td>
<td>$T_j$</td>
</tr>
</tbody>
</table>

Table 1.2

<table>
<thead>
<tr>
<th>Score on Studied Item</th>
<th>1</th>
<th>0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>$p_{Rj}$</td>
<td>$q_{Rj}$</td>
<td>1</td>
</tr>
<tr>
<td>Focal</td>
<td>$p_{Fj}$</td>
<td>$q_{Fj}$</td>
<td>1</td>
</tr>
</tbody>
</table>

Since DIF statistics can be based on the probability of success for the reference and focal groups, the table of raw data can be converted to a table of proportions (see Table 1.2).
If we assume that values of the marginal totals, $N_{Rj}$ and $N_{Fj}$, are fixed and regard the data for the reference and focal groups as having arisen as random samples of size $N_{Rj}$ and $N_{Fj}$, then we can consider $A_j$ and $C_j$ as being distributed as independent binomial variates with parameters $(N_{Rj}, p_{Rj})$ and $(N_{Fj}, p_{Fj})$, respectively. The hypothesis of no DIF corresponds to the null hypothesis

$$H_0: p_{Rj} = p_{Fj} \quad \text{for all } j.$$  \hspace{1cm} (1.7)

Early contingency table methods considered the weighted mean difference between the probabilities of success at each observed score category (Dorans & Kulick, 1983). This procedure, commonly referred to as the standardized $p$-difference ($STDP-DIF$), can be expressed as

$$STDP-DIF = \frac{\sum_{j=1}^{n} w_j (p_{Fj} - p_{Rj})}{\sum_{j=1}^{n} w_j}$$  \hspace{1cm} (1.8)

where $w_j$ is the weight assigned to the difference in probability of correct response at $j$th score category. Generally, the weights are taken to be either the number of focal group members at observed score $j$, or the combined number of reference and focal group members at $j$. Note that $STDP-DIF$ is based theoretically on the difference between the item-test regressions of the focal and reference groups, weighting the differences appropriately. A standard error for $STDP-DIF$ is presented in Dorans and Holland (1993, p. 50).

Related to $STDP-DIF$ is the simultaneous item bias test (SIBTEST) proposed by Shealy & Stout (1993b). While $STDP-DIF$ uses the empirical item-test regression, SIBTEST
regresses item performance onto an estimate of true score based on the Kelley correction (Kelley, 1923, 1947; Lord & Novick, 1968, p. 65), which adjusts the observed tests scores for measurement error. Differences in the empirical item-true score regressions for the focal and reference groups are averaged across score levels with a focal group weighting function. The true score correction improves the matching variable in a way that leads to unbiased estimation of the DIF index when group target ability distributions differ.

The most popular of the contingency table DIF statistics is the Mantel-Haenszel (MH) procedure, originally developed for use in epidemiological research by Mantel and Haenszel (1959), and first applied to DIF detection by Holland and Thayer (1988). The MH procedure tests the null hypothesis

\[ H_0: \frac{p_{Rj}}{q_{Rj}} = \frac{p_{Fj}}{q_{Fj}} \quad j = 1, \ldots, n. \]  

(1.9)

against the alternative hypothesis

\[ H_1: \frac{p_{Rj}}{q_{Rj}} = \alpha_j \left( \frac{p_{Fj}}{q_{Fj}} \right) \quad j = 1, \ldots, n \]  

(1.10)

where \( \alpha_j \) is not equal to one. The null case corresponds to the condition in which \( \alpha_j \) equals one. The equality shown in Equation 1.10 can be expressed as

\[ \alpha_j = \left( \frac{p_{Rj}}{q_{Rj}} \right) \left( \frac{p_{Fj}}{q_{Fj}} \right) = \frac{p_{Rj}q_{Fj}}{p_{Fj}q_{Rj}} \quad j = 1, \ldots, n \]  

(1.11)

indicating that the parameter \( \alpha \) is the odds ratio for the \( n \) two-by-two tables. Thus, the Mantel-Haenszel null hypothesis can be stated as the event to the odds ratio at each score
category equals one.

An estimate of $\alpha$ across all score categories is derived by pooling the odds ratios of the $n$ two-by-two tables using

$$\alpha_{MH} = \frac{\sum A_j D_j / T_j}{\sum B_j C_j / T_j}.$$  \hspace{1cm} (1.12)

Although not immediately apparent by Equation 1.12, the weight assigned to each score level is on the order of

$$\frac{N_{R_j} N_{F_j}}{N_{R_j} + N_{F_j}}$$ \hspace{1cm} (1.13)

which is a measure of the total between-group information at that score level (Mantel & Haenszel, 1959, p. 732).

The odds ratio presented in Equation 1.12, known as the common odds ratio, is on the scale zero to infinity, with $\alpha_{MH} = 1$ corresponding to no DIF. The value of $\alpha_{MH}$ is the average factor by which the odds that a member of the reference group is correct on the studied item exceeds the corresponding odds for a comparable member of the focal group. Values of $\alpha_{MH}$ will exceed 1 for items on which the reference group performed better than did comparable members of the focal group.

The common odds ratio can be transformed by the natural logarithm to give the Mantel-Haenszel log-odds-ratio ($MHLOR$)

$$MHLOR = \ln (\alpha_{MH}).$$ \hspace{1cm} (1.14)

It is convenient to transform the $MHLOR$ to a symmetric scale in which zero is the null value.
Such a scale is the $MHD-DIF$ index which can be expressed as

$$MHD-DIF = (-4/1.7) MHLOR = -2.35 MHLOR.$$  \hspace{1cm} (1.15)

The $MHD-DIF$ index is based on a conversion of the odds ratio to a difference in the reference and focal group values on the delta metric, which has a mean of 13 and a standard deviation of 4 (Dorans & Holland, 1993). One unit on the $MHD-DIF$ scale is approximately equal to a difference of ten points in the percentage of correct responses by members of the two groups. The $MHD-DIF$ index is negative for DIF against the focal group, positive for DIF against the reference group, and zero in the absence of DIF.

Mantel and Haenszel (1959) also offered a chi-square test statistic with one degree of freedom that can be used to perform a statistical test of uniform DIF (see Camilli & Shepard, 1994, p. 120). The chi-square statistic is given by

$$MH\chi^2 = \frac{\left\{ \sum_{j=1}^{n} \left[ A_j - E(A_j) \right] - \frac{1}{2} \right\}^2}{\sum_{j=1}^{n} \text{VAR}(A_j)} \hspace{1cm} (1.16)$$

where

$$\text{VAR}(A_j) = \frac{N_{Rj} N_{Fj} N_{1j} N_{0j}}{T_j^2 (T_j - 1)}$$ \hspace{1cm} (1.17)

and

$$E(A_j) = \frac{N_{Rj} N_{1j}}{T_j}.$$ \hspace{1cm} (1.18)

The effectiveness of the MH procedure to detect DIF has resulted in its frequent use
in bias analyses of the items used in large scale testing programs. Going beyond the
dichotomous significant/non-significant \( M_{H}X^{2} \) result, Educational Testing Service (ETS) has
developed a system for flagging items by classifying them into one of three categories
according to the magnitude of the associated \( MHD-DIF \) (Zieky, 1993). The definitions of the
categories are as follow:

Category A) \( MHD-DIF \) not significantly different from zero, or absolute value less
than 1.0.

Category B) \( MHD-DIF \) significantly different from zero and absolute value of at
least 1.0, and either: i) less than 1.5, or ii) not significantly greater than
1.0.

Category C) \( MHD-DIF \) significantly greater than 1.0 and absolute value greater
than or equal to 1.5.

The level of alpha used in the tests of significance is not stated in the description given by
Zieky (1993), but is assumed to be 0.05. Items falling in Category A are considered to have
negligible DIF, those in Category B are considered to have slight to moderate DIF, and those
in Category C are considered to have moderate to high DIF.

**Definition of the Problem: Contamination-Induced Misclassification**

Let there exist two populations, termed reference and focal, each having a unique
distribution of target ability \( (\theta) \). For a given item, let the probability of correct response for
each population be defined by an ICC unique to that population. We have for each
population a bivariate distribution of total test score \( (m) \) and \( \theta \), where total test score is a
discrete random variable and \( \theta \) is a continuous random variable. Thus, after conditioning on the \( j \)th level of total test score, for each population there exists a distribution of target ability values, represented by

\[
f_R(\theta | m = j)
\]

for the reference population, and by

\[
f_F(\theta | m = j)
\]

for the focal population.

This thesis is concerned with the effects of using the total test score as a means to match reference and focal population members of equal ability on the performance of the MH procedure to detect DIF when the test contains one or more biased items. When the test contains items that are biased against one of the populations, the test score is not uniquely determined by target ability. As a result of this invalidity, the test score is said to be contaminated, where contamination is defined as an unintended multidimensionality of the test score such that examinees of one group are misclassified to test score categories that are different from those expected had the test score been uniquely determined by target ability. Misclassification caused by test score contamination is referred to here as contamination-induced misclassification (CIM). A consequence of contamination is that

\[
f_R(\theta | m = j) \neq f_F(\theta | m = j)
\]

(1.19)

for all \( j \).

Let us assume that for a particular test all items containing bias are biased against the focal population. Then there exist focal population members who are systematically misclassified to test score levels below the levels expected had no bias existed, causing the
distribution of target ability for the focal population at total test score $m = j$ to be shifted to the right of that of the reference population at $m = j$. Under this situation, the following relationship is assumed to hold

$$E(\theta | G = F, m = j) > E(\theta | G = R, m = j)$$  \hspace{1cm} (1.20)

where $G$ denotes group membership. Using the relationship defined in Equation 1.20, the following result holds.

**Result 1.1:** When the studied item functions identically for the reference and focal groups, and CIM exists, the probability of correct response for the focal population having observed score $j$ will exceed that for the reference population having observed score $j$.

**Proof:** Assuming the function relating probability of correct response to target ability is monotonic increasing, considering the effects of contamination presented in Equation 1.20 gives

$$E(Y|G = F, m = j) > E(Y|G = R, m = j).$$  \hspace{1cm} (1.21a)

or equivalently

$$E(Y|G = F, m = j) = E(Y|G = R, m = j) + k$$  \hspace{1cm} (1.21b)

where $k$ is some constant greater than zero.

The effects of contamination on $\alpha_{MH}$ as a test of the null hypothesis that the odds ratio equals unity is stated in Result 1.2.

**Result 1.2:** In the presence of CIM, $\alpha_{MH}$ is an incorrect test of the null hypothesis that the odds ratio equals unity.
**Proof:** Let us consider the null case in which the ICC for the studied item is identical for the reference and focal populations. If we consider $\alpha_{ali}$ as a weighted composite of the odds ratios over all $n$ matching categories, where all weights are positive, then it suffices to show that, in the presence of CIM, the expected odds ratio for score category $j$ will exceed the null value of unity even when no bias exists. Let the probability of correct and incorrect response in the population be denoted by $p$ and $q$, and the reference and focal populations by $R$ and $F$. Using Result 1.1, which showed that in the presence of CIM $p_{R} = p_{F} + k$ when focal and reference ICCs are identical for the studied item and CIM exists, the odds ratio for the $j$th matched set of reference and focal group members can be expressed as

$$
\alpha_{j} = \frac{p_{Rj}q_{Fj}}{q_{Rj}p_{Fj}}
$$

(1.22a)

$$
= \frac{p_{Rj}(q_{Rj} - k)}{q_{Rj}(p_{Rj} + k)}
$$

(1.22b)

$$
= \frac{p_{Rj}q_{Rj}}{q_{Rj}(p_{Rj} + k)} - \frac{p_{Rj}k}{q_{Rj}(p_{Rj} + k)}
$$

(1.22c)

$$
= \frac{q_{Rj}}{q_{Rj} + k} - \frac{p_{Rj}k}{p_{Rj} + k}
$$

(1.22d)

$$
= 1 - \frac{k}{(p_{Rj} + k)} + \frac{p_{Rj}k}{q_{Rj}(p_{Rj} + k)}
$$

(1.22e)
\[ 1 - \frac{q_{Rj}k + p_{Rj}k}{q_{Rj}(p_{Rj} + k)} \]  
\[ = 1 - \frac{(q_{Rj} + p_{Rj})k}{q_{Rj}(p_{Rj} + k)} \]  
\[ = 1 - \frac{k}{q_{Rj}(p_{Rj} + k)} \].

Since it is assumed that \( k \) is some constant greater than zero, and the expected values of \( p \) and \( q \) are nonzero by definition, it follows that the value of the odds ratio will be less than unity in the presence of CIM, even when the ICC’s for the focal and reference populations are identical. Since \( \alpha_{MH} \) is a weighted composite of the odds ratios over all \( n \) matching categories, where all weights are positive, it follows that in the presence of CIM the expected value of \( \alpha_{MH} \) is less than unity, and thus \( \alpha_{MH} \) is an incorrect test of the null DIF hypothesis that the odds ratio is unity. 

The findings of result 1.2 imply that when test score contamination caused by items biased against the focal group exists, the following two outcomes hold: (1) under the condition of no bias in the studied item, \( \alpha_{MH} \) is expected to be less than unity, and (2) under the condition of bias against the focal group for the studied item, \( \alpha_{MH} \) is expected to underestimate the population odds ratio obtained when reference and focal group members are matched on a valid measure of ability.
Practical Implications of CIM on DIF Detection

The primary implication of CIM is that the MH null hypothesis may not be rejected, even when substantial DIF exists. Furthermore, this loss of power of the MH procedure to detect an odds ratio differing from unity becomes more severe when several of the non-studied items contain bias and thus CIM aggregates over these items. The most severe case in applied testing situations would be when non-studied items contain moderate levels of bias, since they would not have great enough levels of DIF to warrant removal from the test, but would contain enough bias to cause some amount of CIM. In a situation where cutoff points are used to determine whether or not to flag an item (for example, the ETS classification system), the presence of aggregated CIM could have serious implications for DIF detection. Not only could it cause a given item to go undetected, but it could also lead to a situation in which the probability of detection is contingent on the properties of the non-studied test items.

How much could CIM influence DIF detection on a typical test? While a precise estimate of the percentage of items that normally contain DIF is not available, several researchers have shown that in actual tests over 20% of the items contained significant levels of DIF (Hambleton & Rogers, 1989; Mazor, Kanjee & Clauser, 1993; Oshima & Miller, 1992). With CIM possibly aggregating over 20% of the items, it is easy to see that their combined effects could lead to a situation whereby some focal group examinees are placed at observed score categories one, two, three or more score categories below their expected placement in the absence of bias. Applying these results to the effects of contamination depends on the extent to which DIF is representative of bias, as well as the extent to which
the DIF consistently acts against the same group.

CIM also has implications for interpreting the results of simulation studies investigating the properties of DIF statistics under varying conditions. These simulation studies are often conducted using a simulated test in which up to 20% of the items contain large levels of DIF (see for example Swaminathan & Rogers, 1990). As a consequence, the results of such simulation studies may be biased due to the affects of CIM on the DIF statistics under investigation.

Previous Studies of the Effects of CIM

The possible effects of matching criterion contamination have received a moderate amount of attention in the DIF detection literature. Clauser, Mazor and Hambleton (1993) found that eliminating contamination led to an increase in the detection of DIF in the studied item using the $MH \chi^2$, suggesting that contamination can have profound effects on interpreting the level of DIF in an item. This result was supported by Shealy and Stout (1993b), who showed that the mean absolute value of $MHD-DIF$ decreased slightly as contamination increased. This result was not supported, however, by Donoghue, Holland and Thayer (1993), who found that contamination did not significantly affect the absolute value of the $MHD-DIF$. This inconsistency is likely due to differences in the magnitude of contamination introduced into the matching criterion. While Donoghue, Holland and Thayer (1993) introduced slight to moderate levels of contamination into the non-studied items (up to 10% of the items containing an increase in the difficulty parameter of 0.3 for the focal group), Clauser, Mazor and Hambleton (1993) introduced higher levels of contamination into the
matching criterion (up to 20% of the items containing an increase in the difficulty parameter of 0.6 for the focal group). These conflicting findings suggest that the issue of contamination deserves further consideration. Such additional research should quantify the effects of contamination on: (1) the power and Type I error rate of $MH\chi^2$, (2) the mean value of the $MHD-DIF$, and (3) the probability of flagging an item for removal using the ETS classification system. If it is found that contamination substantially affects the MH procedure, then methodology should be developed to adjust the MH procedure for such effects.

**Organization and Objectives**

This thesis was designed to investigate the problem of CIM by addressing the following three questions:

1. To what extent can CIM be expected to affect the results of the MH procedure?
2. If CIM does affect the MH procedure, is it possible to adjust the MH value to correct for such effects?
3. Is it possible to develop a DIF detection procedure that is robust to the effects of CIM?

To this end, the core of this dissertation has been organized into seven chapters, each of which addresses a distinct aspect of the investigation as outlined above. The purpose, content and methodology used in each of these chapters are as follows:

Chapter Two is the report of an empirical investigation quantifying the effects of CIM on the performance of the MH procedure to detect DIF. By means of a simulation study, an
examination is made of the effects of CIM on the power and Type I error rate of the MH procedure. To determine how CIM affects bias detection in practical testing situations, the effect of CIM is also assessed in terms of its influence on the ETS classification of items.

In Chapters Three to Five, I describe the development of an adjustment to the MH procedure that corrects for the effects of CIM, and empirically test the performance of the adjustment. The adjustment is first derived theoretically in Chapter Three, followed by a description of the procedures used to estimate the parameters of the adjustment in Chapter Four. Chapter Five assesses the performance of the adjustment using a simulation study.

Chapter Six contains a proposal for a new DIF detection statistic, referred to as MB-DIF, which is hypothesized to have the advantage of being robust to the effects of CIM. Chapter Seven investigates the performance of MB-DIF using a simulation study which assesses the power and Type I error rates of MB-DIF under varying levels of CIM and sample size.

A final chapter discusses the practical implications and limitations of the results of this dissertation.
Chapter Two

Assessing the Effects of CIM on the MH Procedure

Introduction

It was proven in Chapter One that CIM causes $\alpha_{MH}$ to deviate from the null hypothesis value of unity when the studied item functions identically for the reference and focal populations (see Result 1.2). It was also shown that the direction of the deviation from unity is dependent on the direction of the item bias causing the contamination; when contamination is caused by bias against the focal group, $\alpha_{MH}$ is less than unity, indicating DIF against the reference group. This result can be generalized to the case where the studied item is biased against the focal group, such that as the magnitude of bias against the focal group across multiple test items increases, the power of the MH procedure to detect DIF in the studied item against the focal group decreases. The precise magnitude and practical implications of the effects of CIM on the performance of the MH procedure in assessing DIF remains unknown. This problem is addressed through a consideration of three issues: (1) the extent to which the MH measurement of DIF in a given item is contingent upon the magnitude of CIM, (2) the extent to which CIM affects the classification of items using the ETS classification system (see Chapter One for a description of this system), and (3) the extent to which CIM causes the MH procedure to indicate the presence of negative DIF (corresponding to bias against the reference group) in items containing no bias.
Method

The simulations presented below were based on an artificial test consisting of dichotomously scored items. The parameters of the artificial items were those of a three parameter logistic regression model (3PL). For each item, the difficulty parameter \( b \) was drawn from a normal distribution with a mean of zero and standard deviation of one. Item discrimination parameters \( a \) were sampled from a log-normal distribution where \( a \) is taken as the exponent of \( z \), and where \( z \) is a normal deviate with a mean of zero and a standard deviation of 0.1225. These parameter distributions are the same as those used in previous research, and represent realistic distributions of item parameters (see Donoghue & Allen, 1993). All items were assigned a \( c \)-parameter value of 0.2.

Generation of the simulated test data was conducted by: 1) drawing a standard normal variate \( \theta \), 2) computing the probability of success \( P \) on an item for \( \theta \) using the item’s 3PL, 3) drawing a uniform deviate \( U \), and 4) setting the item response equal to 0 if \( U > P \) and 1 for \( U \leq P \). DIF was introduced by increasing the item’s \( b \)-parameter for the focal group by a constant \( t \), making the item more difficult for the focal group relative to the reference group.

For each of the conditions listed below, \( MHD-DIF \) and \( MHX^2 \) were computed. The effect of CIM was assessed by comparing across varying levels of contamination the mean values of \( MHD-DIF \), the proportion of items having statistically significant \( MHX^2 \) at a 0.05 significance level, and the proportion of items classified as ETS Category C.

For all simulations, the number of examinees in the focal and reference groups were each equal to 1000. The five factors examined in this study were: (1) the number of items on
the test, (2) the number of contaminating non-studied items, (3) the magnitude of DIF induced in the contaminating items, (4) the magnitude of DIF induced in the studied item, and (5) the equality of the means of the focal and reference group ability distributions. Each of these factors is discussed below.

**Factor 1: Number of Test Items.** Simulations were conducted using tests of 20 and 60 items. The rationale for these test lengths was to permit the examination of the effects of CIM on a multiple choice test of typical length (60 items), as well as on a relatively short test, such as might be used in a classroom (20 items). These test lengths are consistent with those used in previous simulation studies investigating the properties of the MH statistic under varying conditions (Clauser, Mazore & Hambleton, 1993).

**Factor 2: Number of Contaminating Items.** Simulations were conducted using an artificial test in which approximately 5%, 10%, and 20% of the non-studied items contained DIF. For tests containing 20 items, the 5%, 10%, and 20% corresponded to 1, 2, and 4 non-studied items containing DIF. For tests containing 60 items, the 5%, 10%, and 20% corresponded to 3, 6, and 12 non-studied items containing DIF. The effect of CIM caused by bias in the studied item was not investigated here for two reasons. First, the effects of CIM produced by the studied item are of less practical importance than those produced by non-studied items. This can be explained as follows. We are only concerned with the effects of CIM on items having moderate-to-high levels of bias, since it is only these items that run the risk of being incorrectly declared DIF-free. While CIM arising from bias in the studied item will affect the testing of DIF in that item, the degree to which the assessment of DIF is affected is fairly equal for all studied items having moderate-to-high levels of bias, and thus would not greatly

25
affect the overall ordering of the magnitude of DIF across such items. That is, contamination associated with the studied items will affect absolute DIF decisions, but not relative DIF decisions. Second, the strategy used here to assess the effects of CIM on the MH procedure was to compare the MH values obtained under varying levels of CIM to a control condition in which no contamination existed. Unfortunately, there is no empirical way of obtaining a "true" control condition in which the studied item has DIF and there is no CIM. As a consequence, it is impossible to control for the CIM caused by the studied item, and thus the effects of CIM caused by the studied item are not addressed here.

Factor 3: Magnitude of DIF in Contaminating Items. Three levels of DIF were considered for the non-studied items; \( t = 0.2, 0.4, \) and \( 0.6 \). Pilot simulations were conducted in the search for values of \( t \) that, in the presence of varying levels of CIM, would result in items being categorized as A, B, and C according to the ETS classification system. These pilot simulations showed that for sample sizes of 1000, \( t = 0.2 \) generally led to items categorized as A, \( t = 0.4 \) to items categorized as B, and \( t = 0.6 \) to items categorized as B or C. As a consequence, these values of \( t \) generate the range of DIF values found in typical testing situations.

Factor 4: Magnitude of DIF in the Studied Item. Two levels of DIF were considered in the studied item; \( t = 0.0 \) and \( 0.6 \). As discussed above, the value of \( t = 0.6 \) lead to DIF values that consistently placed items in ETS categories B and C in the presence of moderate contamination. Since the goal of this chapter was to quantify the effects of CIM on the MH procedure, and also to show the practical implications of CIM on DIF detection, the level of DIF in the studied item was intentionally set to a level that would make its misclassification a
possible result. This possibility is greatest for DIF levels that are relatively close to the threshold between categories B and C. The additional condition of $t = 0.0$ permits the investigation of the effects of CIM on MHD-DIF when no DIF is introduced in the studied item.

*Factor 5: Equality of Group Ability Means.* Consideration was given to two levels of difference in the mean of the focal and reference group ability distributions. The first level was a zero difference between the means of the group ability distributions ($\mu_R = \mu_F = 0.0$). The second level placed the mean focal group ability distribution one standard deviation below that of the reference group ($\mu_R = 0.0$, $\mu_F = -1.0$).

All conditions in which non-studied items contained DIF were crossed, giving a total of 72 conditions (2 levels of test length $\times$ 3 levels of the number of contaminating items $\times$ 3 levels of DIF in non-studied items $\times$ 2 levels of DIF in studied item $\times$ 2 levels of difference in group mean ability). The number of examinees in each group was 1000. For each condition, 1000 replications were run. An additional set of four conditions (one for each of the four possible combinations of test length and equality of group ability distribution means) with $t = 0.0$ for all non-studied items was included to establish baseline results.

**Results**

Table 2.1 displays the mean values of the MHD-DIF for the 20-item test containing a studied item having simulated DIF of $t = 0.6$ across all levels of contamination. The top row of Table 2.1 represents the control condition in which no contamination was introduced from non-studied items. Consider first the left side of Table 2.1, displaying results for the
condition of equal ability distribution means. The results suggest that as contamination increases, the mean value of $MHD$-$DIF$ decreases relative to the value obtained in the control condition of zero contamination. The decrease in mean $MHD$-$DIF$ value was virtually zero under slight contamination (one non-studied item containing moderate DIF), but increased as the level of contamination increased. Under the most extreme levels of contamination simulated here (four items containing large DIF), the underestimation reached 0.303. This condition was also associated with a decrease of 0.04 in the proportion of items deemed to contain significant DIF using $MH\chi^2$, and a decrease from 0.47 to 0.19 in the proportion of items placed in Category C of the ETS categorization scheme. This decrease in the proportion of items flagged as Category C represents a 60% decrease in the proportion obtained in the control condition. The effects of CIM on the mean $MHD$-$DIF$ value and probability of a Category C classification for the conditions in which the magnitude of DIF in the studied item is 0.6 are displayed graphically in Figure 2.1.

The mean $MHD$-$DIF$ for the 20-item test when the means of the ability distributions of the focal and reference groups differ by one standard deviation are displayed in the right side of Table 2.1. In general, when the means of the group ability distributions are different, the effect of CIM had the same trend as when the means were equal. This is shown graphically in Figure 2.1. The effects were, however, less severe than when the ability means were equal, leading to a maximal decrease in the mean value of $MHD$-$DIF$ of 0.228 and a maximal decrease in the proportion of trials having Category C levels of DIF of 0.13. This result suggests that the effects of contamination are more serious in the case of equal group ability means than unequal group ability means. Possible reasons for this result are proposed
in the discussion section.

Table 2.1
Simulation Results for the Conditions in which DIF is Introduced in the Studied Item (t = 0.6), Test Length is 20 Items

<table>
<thead>
<tr>
<th>%</th>
<th>t</th>
<th>MHD-DIF</th>
<th>Power</th>
<th>C</th>
<th>MHD-DIF</th>
<th>Power</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>-1.461</td>
<td>0.95</td>
<td>0.47</td>
<td>-1.178</td>
<td>0.79</td>
<td>0.33</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>-1.463</td>
<td>0.95</td>
<td>0.45</td>
<td>-1.181</td>
<td>0.79</td>
<td>0.32</td>
</tr>
<tr>
<td>5</td>
<td>0.4</td>
<td>-1.458</td>
<td>0.96</td>
<td>0.46</td>
<td>-1.133</td>
<td>0.77</td>
<td>0.32</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>-1.395</td>
<td>0.94</td>
<td>0.41</td>
<td>-1.139</td>
<td>0.78</td>
<td>0.30</td>
</tr>
<tr>
<td>10</td>
<td>0.2</td>
<td>-1.412</td>
<td>0.95</td>
<td>0.41</td>
<td>-1.159</td>
<td>0.79</td>
<td>0.31</td>
</tr>
<tr>
<td>10</td>
<td>0.4</td>
<td>-1.385</td>
<td>0.95</td>
<td>0.39</td>
<td>1.103</td>
<td>0.76</td>
<td>0.30</td>
</tr>
<tr>
<td>10</td>
<td>0.6</td>
<td>-1.331</td>
<td>0.93</td>
<td>0.33</td>
<td>-1.039</td>
<td>0.72</td>
<td>0.27</td>
</tr>
<tr>
<td>20</td>
<td>0.2</td>
<td>-1.380</td>
<td>0.94</td>
<td>0.39</td>
<td>-1.115</td>
<td>0.76</td>
<td>0.31</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>-1.286</td>
<td>0.93</td>
<td>0.29</td>
<td>-0.997</td>
<td>0.73</td>
<td>0.21</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>-1.158</td>
<td>0.91</td>
<td>0.19</td>
<td>-0.950</td>
<td>0.70</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Note. % refers to the percentage of non-studied items containing DIF, t refers to the level of DIF introduced in the non-studied items, MHD-DIF is the mean value of the MHD-DIF over 1000 replications, Power is the proportion of replications having a significant value of the $MHX^2$, and C refers to the proportion of replications having items categorized as C in the ETS scheme.
Figure 2.1
The Effects of Contamination on DIF Detection

20 Item Test

![Graph showing mean MHD-DIF and proportion of Category C items as a function of non-studied item DIF.](image)

60 Item Test

![Graph showing mean MHD-DIF and proportion of Category C items as a function of non-studied item DIF.](image)

Figure 2.1. The mean absolute value of $MHD$-$DIF$ and proportion of Category C items are displayed as a function of non-studied item DIF ($t = 0.2, 0.4, 0.6$). In each graph six conditions are displayed representing each combination of 5%, 10% and 20% of non-studied items containing DIF, and equal (E) and unequal (U) means of the group ability distributions. Standard errors are on the order of 0.008 for the mean $MHD$-$DIF$ values, and 0.01 for the proportion of Category C items.
Table 2.2 shows the mean $MHD-DIF$ values of the 60-item test for the conditions of equal and unequal ability distribution means. The effects of contamination were consistent with those for the 20-item case (see Figure 2.1). It is interesting to note that across both conditions of mean ability differences (equal and unequal), the mean $MHD-DIF$ value was generally larger in absolute value for the 60-item test than for the 20-item test (see Figure 2.1).

Table 2.2
Simulation Results for the Conditions in which DIF is Introduced in the Studied Item ($t = 0.6$), Test Length is 60 Items

<table>
<thead>
<tr>
<th>%</th>
<th>$t$</th>
<th>$MHD-DIF$</th>
<th>Power</th>
<th>$C$</th>
<th>$MHD-DIF$</th>
<th>Power</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>-1.502</td>
<td>0.96</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>-1.488</td>
<td>0.97</td>
<td>0.49</td>
<td>-1.236</td>
<td>0.84</td>
<td>0.34</td>
</tr>
<tr>
<td>5</td>
<td>0.4</td>
<td>-1.452</td>
<td>0.96</td>
<td>0.47</td>
<td>-1.187</td>
<td>0.82</td>
<td>0.33</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>-1.424</td>
<td>0.95</td>
<td>0.45</td>
<td>-1.179</td>
<td>0.82</td>
<td>0.33</td>
</tr>
<tr>
<td>10</td>
<td>0.2</td>
<td>-1.451</td>
<td>0.96</td>
<td>0.48</td>
<td>-1.220</td>
<td>0.83</td>
<td>0.34</td>
</tr>
<tr>
<td>10</td>
<td>0.4</td>
<td>-1.412</td>
<td>0.96</td>
<td>0.42</td>
<td>-1.151</td>
<td>0.82</td>
<td>0.30</td>
</tr>
<tr>
<td>10</td>
<td>0.6</td>
<td>-1.329</td>
<td>0.98</td>
<td>0.37</td>
<td>-1.124</td>
<td>0.79</td>
<td>0.27</td>
</tr>
<tr>
<td>20</td>
<td>0.2</td>
<td>-1.400</td>
<td>0.95</td>
<td>0.42</td>
<td>-1.108</td>
<td>0.80</td>
<td>0.27</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>-1.276</td>
<td>0.93</td>
<td>0.32</td>
<td>-1.057</td>
<td>0.79</td>
<td>0.24</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>-1.169</td>
<td>0.92</td>
<td>0.22</td>
<td>-1.005</td>
<td>0.77</td>
<td>0.20</td>
</tr>
</tbody>
</table>

*Note.* See Table 2.1 for a description of headers.
This finding is to be expected because the studied item is a greater proportion of the 20-item test than the 60-item test, and as a result the contamination associated with the studied item should have a greater effect on the 20-item test than on the 60-item test.

As was the case for the 20-item test condition, the effects of CIM on the 60-item test were more pronounced when group ability distribution means were equal than when they were unequal. For example, the underestimation of $MHD-DIF$ was 0.33 under maximal contamination when ability distribution means were equal, but only 0.26 when ability distribution means were unequal. This underestimation was even more pronounced in the proportion of items classified as Category C; a decrease of 0.29 under maximal contamination (a 57% reduction from control levels) in the equal mean ability case, and 0.17 (46% reduction from control levels) in the unequal mean case.

Tables 2.3 and 2.4 display the mean value of $MHD-DIF$ for an item in which no DIF was introduced. These results show that the $MHD-DIF$ values exceeded the null values of zero, and reached approximately 0.30 under high levels of contamination when the means of the ability distributions were equal for the reference and focal groups. This positive bias of the $MHD-DIF$ is less severe when group ability distributions had unequal means, reaching approximately 0.23 for both test lengths. This finding suggests that the MH procedure can indicate considerable negative DIF even when no DIF exists in the item. This result supports the possibility that CIM can generate negative DIF. This claim is also supported by the increase in Type I error rates displayed in Tables 2.3 and 2.4, which soared to 0.21 for the 20-item test under the largest levels of contamination.
Table 2.3
Simulation Results for the Conditions in which no DIF is Introduced in the Studied Item, and Test Length is 20 Items

<table>
<thead>
<tr>
<th>%</th>
<th>$t$</th>
<th>Unequal Means</th>
<th>Equal Means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$MHD$-DIF</td>
<td>Type I</td>
</tr>
<tr>
<td>0</td>
<td>0.0</td>
<td>-0.003</td>
<td>0.12</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>-0.027</td>
<td>0.13</td>
</tr>
<tr>
<td>5</td>
<td>0.4</td>
<td>0.006</td>
<td>0.16</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>0.036</td>
<td>0.16</td>
</tr>
<tr>
<td>10</td>
<td>0.2</td>
<td>0.013</td>
<td>0.12</td>
</tr>
<tr>
<td>10</td>
<td>0.4</td>
<td>0.066</td>
<td>0.13</td>
</tr>
<tr>
<td>10</td>
<td>0.6</td>
<td>0.091</td>
<td>0.15</td>
</tr>
<tr>
<td>20</td>
<td>0.2</td>
<td>0.074</td>
<td>0.14</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.138</td>
<td>0.14</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>0.229</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Note. % refers to the percentage of non-studied items containing DIF, $t$ refers to the level of DIF introduced in the non-studied items, $MHD$-DIF is the mean of the absolute value of the $MHD$-DIF over 1000 replications, and Type I is the proportion of replications having a significant value of the $MH\chi^2$.

Table 2.4
Simulation Results for the Conditions in which no DIF is Introduced in the Studied Item, and Test Length is 60 Items

<table>
<thead>
<tr>
<th>%</th>
<th>$t$</th>
<th>Unequal Means</th>
<th>Equal Means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$MHD$-DIF</td>
<td>Type I</td>
</tr>
<tr>
<td>0</td>
<td>0.0</td>
<td>0.003</td>
<td>0.06</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>-0.002</td>
<td>0.05</td>
</tr>
<tr>
<td>5</td>
<td>0.4</td>
<td>0.017</td>
<td>0.05</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>0.046</td>
<td>0.06</td>
</tr>
<tr>
<td>10</td>
<td>0.2</td>
<td>0.025</td>
<td>0.05</td>
</tr>
<tr>
<td>10</td>
<td>0.4</td>
<td>0.076</td>
<td>0.05</td>
</tr>
<tr>
<td>10</td>
<td>0.6</td>
<td>0.098</td>
<td>0.07</td>
</tr>
<tr>
<td>20</td>
<td>0.2</td>
<td>0.080</td>
<td>0.07</td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.144</td>
<td>0.08</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>0.233</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Note. See Table 2.3 for a description of headers.
It is interesting to note that when group ability distribution means were different, the Type I error rates tended to be larger for the 20-item test than for the 60-item test. This is to be expected since the 60-item total test score is a more reliable measure of ability.

**Discussion**

These results suggest that contamination of the matching criterion leads to substantial loss of power of the MH procedure in detecting DIF. When the group ability distributions had equal means, the largest levels of contamination led to a decrease in the mean $MHD-DIF$ value of over 0.30, with an associated 60% decrease in the probability of flagging the item as Category C. The effects of contamination were less severe when group ability distributions had unequal means, evidenced by smaller effects on the mean value of $MHD-DIF$, the power of $MH\chi^2$, and the proportion of items identified as Category C. At a given proportion of total test items containing contamination, the loss of power was consistent across tests of 20 and 60 items, suggesting that it is the proportion of contaminated test items rather than the actual number of contaminated items that is directly related to the underestimation of $MHD-DIF$. When contamination was present in the absence of DIF in the studied item, the studied item displayed negative DIF, and Type I error rates of $MH\chi^2$ increased to unacceptably high levels. These results suggest that CIM can have substantial consequences in tests containing moderate to high levels of contamination, and thus should be considered when performing DIF analyses and conducting simulation studies in which DIF is introduced into several items of the simulated test.

A primary implication of these results is that the magnitude of DIF measured in a
given item is contingent on the psychometric properties of the non-studied items. This suggests that the same item may measure different levels of DIF if contained in tests of different items, even if the examinee population is identical. This finding is particularly important when an item is contained in an item bank and thus may be included in tests composed of different collections of items. This issue extends to the case of computer adaptive tests, where the attempt is made to measure DIF in items that have been administered in different sets of items to different samples of examinees.

In interpreting these results, readers should note several points. First, CIM poses the greatest danger to items that have moderate to high levels of DIF, since it is for these items that decisions concerning inclusion and exclusion in the test will be most affected by CIM. For this reason, the present study was focused on items having levels of DIF that would place them at Category B or C. The results reported here will not generalize to items having excessively high or low values of DIF.

It should also be noted that the information contained in the percentage of items having a significant value of $MH\chi^2$ is limited given the levels of DIF in the studied item chosen for this study. That is, the percentage of items with significant DIF was so high even under the null condition that it allowed for little change as contamination increased. Results not reported here showed that when $t = 0.4$, the probability of significance fell across all conditions, allowing for a greater change in the percentage of items with significant DIF (at times exceeding 30%) as contamination increased. This result is to be expected since simulating DIF with $t = 0.4$ places the level of DIF in most items much closer to the critical chi-square value needed for statistical significance.
Researchers conducting DIF analyses have dealt with the issue of contamination by purifying the matching criterion through an iterative process, whereby a preliminary MH procedure is conducted using the total test score as the matching criterion. All items flagged as having substantial DIF are then removed from the matching criterion (with the exception of the studied item), and the MH procedure is repeated using the purified matching criterion. Several methods can be used to flag items for removal: (1) the ETS categorization scheme whereby, for example, all Category C items are flagged; and (2) the significance of the $MH\chi^2$ (see Clauser, Mazor & Hambleton, 1993). One possible criticism of the present study is that if the matching criterion is purified by removing all items deemed to have excessive DIF, then much of the contamination simulated in this study would have been eliminated. In response to this criticism, I contend that the contamination levels used here would usually not lead to the contaminating items being removed from the tests. For example, the most highly contaminated items were those for which $t = 0.60$. From the results of this study, when $t = 0.60$, an item has only a 20%-40% chance of being flagged as Category C when there are moderate to high levels of contamination in the test and when ability distribution means are equal. When ability distributions are not equal, this percentage slips to between 20% and 30%. This result was obtained for relatively large sample sizes ($N = 1000$), so the percentage is likely to be even lower for smaller sample sizes. Although the use of $t = 0.6$ for the non-studied items represent an extreme case, it seems consistent with situations found in applied test development. Note that a limitation to the results of this study is that the power of the $MH\chi^2$ was not investigated for the 0.01 level of significance, and thus we don’t know the extent to which a purification procedure based on this level of significance of $MH\chi^2$, as was
used by Clauser, Mazor and Hambleton, (1993), would affect the results presented here.

The results suggest that in the presence of unequal group ability means, the effects of contamination were moderated relative to the case of equal group ability means. This finding has several possible explanations. First, it has been shown that when group ability distributions have unequal means, the MH procedure may be positively or negatively biased, depending on the value of the item discrimination parameter (Zwick, 1990). This effect of the item discrimination parameter may have been to decrease the effect of contamination when group ability distributions had unequal means. A second explanation concerns the effect of ability distribution on DIF detection. A recent study (Penfield, 1999) showed that the performance of the $MH^2$ in detecting DIF is highly contingent on the ability level of the sample of examinees tested. For a given test, as examinee ability increases, the estimate of DIF increases. This effect occurs because the potential for DIF to be expressed is higher in groups of higher ability. Having ability distributions with means differing by one standard deviation (as in this study) in effect shifts the focal group ability distribution to the left by one standard deviation, and thus decreases potential for DIF.

The results of this chapter show that as contamination increases, the magnitude of negative DIF observed in items in which no DIF is simulated increases dramatically. That is, CIM causes negative DIF. This finding raises the issue in applied DIF analyses of the interpretation of negative DIF. It is often the case that many items display slight to moderate levels of negative DIF (Camilli, 1993). This finding is well known, and anticipated given the positive nature of non-parametric DIF analyses. The current results suggest the need for a quantification of the anticipated negative DIF at a particular level of contamination. Given
the relationship between contamination and negative DIF, the mean level of negative DIF may be a useful index of the overall level of contamination.
Chapter Three

Developing an Adjustment to the MH Procedure to Correct for CIM

Introduction

In Chapter One it was shown that CIM can cause $a_{MH}$ to deviate from the null hypothesis value of unity even when the item functions identically for the reference and focal populations (see Result 1.2). Chapter Two added empirical evidence to this result through a simulation study, which showed that when items containing DIF are included in the internal measure of ability, the observed value of $MHD$-DIF decreased from its control value, as did the power of $MH\chi^2$ to detect DIF. This decrease was negligible when only 5% of the items contained small levels of DIF. When 10% or more of the items contained moderate to large levels of DIF, however, the decrease in the $MHD$-DIF value and the power of $MH\chi^2$ became large enough to affect the diagnosis of DIF for the item. These results suggest that matching criterion contamination is an important consideration in the interpretation of the MH procedure, and that investigations into item bias could benefit from a remedy to this problem. This chapter proposes an adjustment to the MH procedure that corrects for the effect of matching criterion contamination.

A Bivariate Latent Trait Model of Item Bias

The adjustment to the MH procedure is grounded theoretically in a bivariate latent trait model of bias. This model includes a definition of item bias that permits the estimation of bias using sample information, and points to several identities that are implicated in
estimating the parameters required for the MH adjustment.

Let a randomly chosen person's observed score on item $i$ ($Y_i$) be a function of two latent variables, $\theta$ and $\eta$, where $\theta$ represents the target ability and $\eta$ represents the nuisance determinant, and each person is assumed to have a fixed value of $\theta$ and $\eta$ at a given testing occasion. Let a correct response on the item be denoted by $Y_i = 1$, and an incorrect response by $Y_i = 0$. The function relating the probability of the event $Y_i = 1$ to $\theta$ and $\eta$ is represented by

$$f(\theta, \eta) = P(Y_i = 1 | \theta, \eta)$$  \hspace{1cm} (3.1)

where the subscript $i$ indicates that the function is specific to item $i$.

Let a given item have threshold values associated with the target ability and nuisance determinant dimensions, such that an examinee must have above threshold values on both dimensions in order to answer the item correctly. Let $\Theta$ be an indicator variable such that $\Theta = 1$ represents an above threshold value on the target ability dimension, and $\Theta = 0$ represents a below threshold value on the target ability dimension. Similarly, let $H$ be an indicator variable such that $H = 1$ represents an above threshold value on the nuisance determinant dimension, and $H = 0$ represents a below threshold value on the nuisance determinant dimension. The respective probabilities of success on each latent dimension are expressed by the functions

$$g_i(\theta) = P(\Theta_i = 1 | \theta)$$  \hspace{1cm} (3.2)

$$h_i(\eta) = P(H_i = 1 | \eta)$$  \hspace{1cm} (3.3)

where the subscript $i$ indicates that the functions are specific to item $i$. It is assumed that for a
fixed value of $\theta$, the function described in Equation 3.2 is constant across all values of $\eta$, and that for a fixed value of $\eta$, the function described in Equation 3.3 is constant across all values of $\theta$.

The model presented here is based on the assumption that the outcome $Y_i = 1$ can occur if and only if the events $\Theta_i = 1$ and $H_i = 1$ are satisfied. Hence, the probability of $Y_i = 1$ conditional on $\theta$ and $\eta$ can be expressed as

$$P(Y_i = 1| \theta, \eta) = P(\Theta_i = 1 \cap H_i = 1|\theta, \eta).$$ (3.4a)

Let us assume that the events $\Theta_i = 1$ and $H_i = 1$ are independent in both the reference and focal populations. It follows that Equation 3.4a can be expressed as

$$P(Y_i = 1| \theta, \eta) = P(\Theta_i = 1|\theta)P(H_i = 1|\eta)$$ (3.4b)

or equivalently,

$$f(\theta, \eta) = g(\theta) h(\eta).$$ (3.4c)

Using this model of the probability that $Y_i = 1$, we can define a model of bias with which to quantify the magnitude of item bias. This model is referred to as manifest bias, where the term manifest is used to distinguish the model of bias presented here from the generic concept of bias.

**Definition 3.1:** The manifest bias for item $i$, $\delta_i(\theta, \eta)$, is defined as

$$\delta_i(\theta, \eta) = g_i(\theta) - f_i(\theta, \eta).$$ (3.5)

Manifest bias has an equivalent interpretation as the product of the probability of success on the target ability dimension and failure on the nuisance determinant dimension, shown by
\[ \delta(\theta, \eta) = g_1(\theta) - f_2(\theta, \eta) \]  
\[ = g_1(\theta) - g_2(\theta) h_3(\eta) \]  
\[ = g_1(\theta)[1 - h_3(\eta)] \]  
\[ = P(\Theta_i = 1|\theta)[1 - P(H_i = 1|\eta)] \]  
\[ = P(\Theta_i = 1|\theta)P(H_i = 0|\eta). \]  

For any fixed value of \( \theta \) and \( \eta \), manifest bias can be interpreted as the probability of incorrectly responding to an item given a sub-threshold value on the nuisance determinant dimension. Thus, for fixed values of \( \theta \) and \( \eta \), as the item threshold for the nuisance determinant dimension increases, the manifest bias can be expected to increase. Examination of Equation 3.6e indicates that manifest bias is bounded between 0 and 1.

**Relevant Assumptions and Results**

The development of the adjustment relies on several assumptions and results.

**Assumption 3.1:** The distribution of \( \eta \) in the reference and focal populations satisfies the following conditions:

\[ P(H_i = 1|\eta, G = R) = 1 \]  
\[ P(H_i = 1|\eta, G = F) \leq 1. \]

**Assumption 3.2:** The function relating the probability of success on the target ability dimension to the value of target ability is identical for the reference and focal groups. That is, if group membership is denoted by \( G \), then

\[ g_1(\theta|G = R) = g_1(\theta|G = F). \]
**Result 3.1:** For the reference group, $f(\theta, \eta)$ is identical to $g(\theta)$ for all values of $\eta$ observed in the reference population.

**Proof:** From Equation 3.4c we know that

$$f(\theta, \eta) = g(\theta)h(\eta).$$

From Assumption 3.1 it is asserted that for all values of $\eta$ observed in the reference population

$$h(\eta) = 1.$$

It follows that for the reference population

$$f(\theta, \eta) = g(\theta)(1) = g(\theta). \quad (3.9)$$

**Result 3.2:** The function defining the probability of observed item success for the reference group is identical to the function defining the probability of success on the target ability dimension for the focal group. That is

$$f(\theta, \eta|G = R) = g(\theta|G = F) \quad (3.10a)$$

or, for any values of $\theta$ and $\eta$

$$P(Y_i = 1|G = R) = P(\Theta_i = 1|G = F). \quad (3.10b)$$

**Proof:** From Assumption 3.2 it is stated that

$$g(\theta|G = R) = g(\theta|G = F),$$

and from Result 3.1 it is known that

$$f(\theta, \eta|G = R) = g(\theta|G = R).$$

Combining the above information we have
Developing the Adjustment

As discussed in Chapter One, the MH procedure assesses the hypothesis that the odds of success are equal for the reference and focal groups at each level of observed score. In particular, \( \alpha_{MH} \) is an overall odds ratio across all score categories, and is obtained by weighting the odds ratio at each score category by the associated between-group information of each score category (Mantel & Haenszel, 1959, p. 732). As shown theoretically in Chapter One (Result 1.2) and empirically in Chapter Two, CIM causes the MH procedure to become an incorrect test of the null hypothesis that the odds ratio is equal to unity. This effect of CIM on the MH procedure is a direct consequence of CIM-induced systematic differences in latent target ability distributions for reference and focal group populations after conditioning on observed score category. Thus, adjusting the MH procedure for the effects of CIM implies adjusting the observed distribution of studied item scores at a given observed score category to the expected value obtained when group ability distributions are not affected by CIM.

Let the overall odds ratio defined under the condition of contamination \( (\alpha_{MH}) \) be distinguished from that under the condition of no contamination \( (\tilde{\alpha}_{MH}) \). Then, \( \tilde{\alpha}_{MH} \) is defined as the overall odds ratio expected when the reference and focal populations at each level of matching category have equal distributions of target ability. The development of the adjustment follows from the following result.
**Result 3.3:** An odds ratio under the condition of no contamination is equal to the observed odds ratio under the condition of contamination for which the observed reference group probability of correct response on the item in question \( P(Y = 1|G = R) \) is replaced with the focal group probability of success on the target ability dimension \( P(\Theta = 1|G = F) \).

**Proof:** The following derivations are conducted for a single hypothetical item, and thus the subscript \( i \) denoting the particular item is dropped without loss of clarity. Let us define an odds ratio under the condition of no contamination as one in which the reference and focal group members being compared have equal target ability distributions. To achieve this equality of target ability distributions for a given group of focal population members, we can assign some arbitrary reference group, denoted \( R_* \), the distribution of \( \theta \) such that it is identical to that of the focal group of interest. That is, if we denote the density of target ability for the focal group of interest by \( f_1(\theta) \), and the density of the target ability for the arbitrary reference group by \( f_2(\theta) \), then by definition

\[
f_1(\theta) = f_2(\theta) \tag{3.11}
\]

The arbitrary reference group described in Equation 3.11 is said to be a validly matched reference group. Using the arbitrary reference group shown in Equation 3.11, the odds ratio obtained under the condition of no contamination for the focal group of interest can be expressed as
Since it is assumed that for the reference group the probability of success is independent of the item threshold for the nuisance determinant dimension, Equation 3.12a can be restated as

\[
\tilde{\alpha} = \frac{\frac{P(\Theta = 1|G = R_A)}{1 - P(\Theta = 1|G = R_A)}}{\frac{P(\Theta = 1 \cap H = 1|G = F)}{1 - P(\Theta = 1 \cap H = 1|G = F)}}.
\] (3.12a)

Using the identity established in Equation 3.11, it follows that Equation 3.12b can be expressed as

\[
\tilde{\alpha} = \frac{\frac{P(\Theta = 1|G = R_A)}{1 - P(\Theta = 1|G = R_A)}}{\frac{P(\Theta = 1 \cap H = 1|G = F)}{1 - P(\Theta = 1 \cap H = 1|G = F)}}.
\] (3.12b)

The top line of Equation 3.12c is equal to the focal group probability of success on the target ability dimension (see Equation 3.2). Thus, to obtain a estimate of the odds ratio obtained under the condition of no contamination for a given focal group, we need only estimate the probability of success on the target ability dimension for the focal group in question.
Let a particular level of the matching criterion be denoted by \( m = j \). Result 3.3 indicated that an adjustment to the MH procedure requires only an estimate of the focal group probability of success on the target ability dimension at each observed score category \( j \).

Using this estimate, the adjusted MH value can be computed by adjusting the observed number of reference group correct and incorrect responses (\( A \) and \( B \) in Table 1.1) such that they match the expected values for a sample of \( N_{rj} \) reference group members with probability of correct response \( P(\Theta = 1|G = F, m = j) \). The adjusted reference group numbers will be represented by \( \tilde{A}_j \) and \( \tilde{B}_j \). This is stated formally in Result 3.4.

**Result 3.4:** The adjusted MH value is obtained by determining the expected number of correct and incorrect responses (\( \tilde{A}_j \) and \( \tilde{B}_j \)) from a sample of \( N_{rj} \) reference group members with probability of correct response \( P(\Theta = 1|G = F, m = j) \). That is

\[
\tilde{A}_j = P(\Theta = 1|G = F, m = j)N_{rj}
\]

\[
\tilde{B}_j = [1 - P(\Theta = 1|G = F, m = j)]N_{rj}.
\]

Using the theory presented in Result 3.4 to estimate the adjusted number of correct and incorrect responses of a validly matched reference group, it is possible to express \( \alpha_{MH} \) obtained under the condition of no contamination as
This adjusted value of $\alpha_{MH}$ can then be appropriately transformed to obtain the adjusted values of $MHD-DIF$ and $MH\chi^2$, as outlined in Equations 1.15 and 1.16. The variance of the $MHLOR$ was derived by Philips and Holland (1987) and Robins, Breslow and Greenland (1986) and is given by

$$S^2 = \frac{\sum_{j=1}^{n} T_j^{-2} (A_j D_j + \alpha_{MH} B_j C_j) [(A_j + D_j + \alpha_{MH} (B_j + C_j) ]}{2 \left( \sum_{j=1}^{n} \frac{A_j D_j}{T_j} \right)^2}. \quad (3.16)$$

The adjusted variance can be expressed as

$$\tilde{S}^2 = \frac{\sum_{j=1}^{n} T_j^{-2} (\tilde{A}_j D_j + \tilde{\alpha}_{MH} \tilde{B}_j C_j) [(\tilde{A}_j + D_j + \tilde{\alpha}_{MH} (\tilde{B}_j + C_j)]}{2 \left( \sum_{j=1}^{n} \frac{\tilde{A}_j D_j}{T_j} \right)^2}. \quad (3.17)$$

Using this estimate of the variance, levels of significance can be assessed for the adjusted values of the $MHLOR$ and $MHD-DIF$. 

\[ \tilde{\alpha}_{MH} = \frac{\sum_{j=1}^{n} (\tilde{A}_j D_j) / T_j}{\sum_{j=1}^{n} (\tilde{B}_j C_j) / T_j}. \]
Chapter Four

Estimating the Parameters of the Adjustment

Introduction

Chapter Three presented the theory used to develop the MH adjustment. This theory culminated in the result (Result 3.3) that the only unknown parameter necessary to the estimation of the adjusted values of \( MHD-DIF \) and \( MH\chi^2 \) is the focal group probability of success on the target ability dimension at observed score category \( j \), represented by \( P(\Theta = 1 \mid G = F, m = j) \). This chapter is dedicated to the estimation of this parameter.

The Estimation Procedures

I begin this chapter with a statistical description of \( P(\Theta = 1 \mid G = F, m = j) \). Since \( \Theta \) is a Bernoulli variable, the probability of success on the target ability dimension for the focal group at observed score category \( m = j \) can be expressed as

\[
P(\Theta = 1 \mid G = F, m = j) = \frac{\sum_{p=1}^{N_{Fj}} (\Theta_p \mid G = F, m = j)}{N_{Fj}}
\]  

(4.1)

where the subscript \( p \) refers to the person \( p \) of a total of \( N_{Fj} \).

The denominator of Equation 4.1 is a known, observed quantity equal to the number of focal group members at observed score category \( j \). The numerator, however, must be estimated. Unfortunately, no simple unbiased estimate of the numerator of Equation 4.1 is available, and thus a series of estimates are required to arrive at the desired quantity. This
chapter outlines this path of estimation, described through six Results. A summary of these procedures is provided at the end of the chapter in Table 4.1.

Let us assume that CIM exists, and define the observed score category, \( m \), by

\[
m = \sum_{i=1}^{n} Y_i
\]  

(4.2)

where \( i \) represents any one of the \( n \) test items. Let us also define a valid score category, \( v \), as the number of target ability dimension successes across all \( n \) test items. That is

\[
v = \sum_{i=1}^{n} \Theta_i .
\]  

(4.3)

Thus each person \( p \) is assumed to have an observed and valid test score, denoted \( m \) and \( v \), respectively. Let a particular level of observed and valid score category be represented by \( j \).

In the presence of CIM, the observed score category \( m = j \) for the focal group is composed of focal group members from valid score categories \( v = j, j + 1, \ldots, j + k \). For example, if there are two items on the test that are biased against the focal group, then the focal group members at observed score category \( m = j \) can be expressed as the combination of those members at observed score \( j \) who belong to valid score categories \( v = j, j + 1, \) and \( j + 2 \). As a result, the numerator of Equation 4.1 has the equivalent expression of

\[
\sum_{p=1}^{N} (\Theta|G = F, m = j) = \sum_{h=0}^{k} \sum_{p=1}^{N|m,v} (\Theta|G = F, v = j + h)
\]  

(4.4)

where \( k \) is the total number of biased items, and \( N|m,v \) is a shortened form of \( (N|m = j, v = j + h) \) which represents the number of focal group members at observed score category \( m = j \) and valid score category \( v = j + h \), where \( h \) equals 0, 1, \ldots, \( k \). Result 4.1 presents the expected
value of the quantity shown in Equation 4.4.

**Result 4.1:** The expected value of the number of successes on the target ability dimension for focal group members with observed score \( m = j \) can be expressed as

\[
E \left[ \sum_{p=1}^{N} (\Theta | G = F, m = j) \right] = E \left[ \sum_{h=0}^{k} \sum_{p=1}^{N|m,v} (\Theta_p | G = F, v = j + h) \right] \quad (4.5a)
\]

\[
= \sum_{h=0}^{k} \sum_{p=1}^{N|m,v} E(\Theta_p | G = F, v = j + h). \quad (4.5b)
\]

There are two comments to be made at this juncture. First, the estimated number of successes on the target ability dimension, as computed using Equation 4.5b, is not likely to be a whole number. Second, two parameters on the right hand side of Equation 4.5b are not observed, and thus must be estimated: \( P(\Theta = 1 | G = F, v = j + h) \), and \( (N|G = F, m = j, v = j + h) \). These estimates are described in Results 4.2 and 4.3.

**Result 4.2:** The mean value of the observed item response for the reference group at observed score category \( m = j \) is an unbiased estimator of the focal population probability of success on the target ability dimension at valid score category \( v = j \). That is

\[ P(\Theta = 1 | G = F, v = j + h) \] is estimated by \( (\bar{Y} | G = R, m = j + h) \). \quad (4.6)

**Proof:** From Assumption 3.2 it is asserted that

\[ g(\theta | G = R) = g(\theta | G = F). \]

Since \( v \) is determined solely by \( \theta \), and if it is assumed that the distribution of \( \theta \) is
equal for the reference and focal populations, then it follows that the distribution of $\theta$ at valid score category $v$ is equal for the reference and focal populations. It follows that

$$P(\Theta = 1 | G = F, v = j + h) = P(\Theta = 1 | G = R, v = j + h). \quad (4.7a)$$

Using Result 3.1 it follows that

$$P(\Theta = 1 | G = R, v = j + h) = P(Y = 1 | G = R, m = j + h) \quad (4.7b)$$

from which it can be stated that

$$P(\Theta = 1 | G = F, v = j + h) = P(Y = 1 | G = R, m = j + h). \quad (4.7c)$$

Result 4.3: The expected number of focal group examinees at observed score category $j$ misclassified from valid score category $v = j + h$, represented by $\left(N \mid G = F, m = j, v = j + h\right)$, can be expressed as the expected value of a binomial distribution with $\left(N \mid G = F, v = j + h\right)$ trials and probability $P(B \mid v = j + h, h)$ of success on each trial, where $(B \mid v = j + h, h)$ represents the event that an examinee from valid score category $v = j + h$ incorrectly responds to $h$ items due to sub-threshold levels of the nuisance determinant dimension. That is,

$$E(N \mid G = F, m = j, v = j + h) = (N \mid G = F, v = j + h)P(B \mid v = j + h, h). \quad (4.8)$$

The right hand side of Equation 4.8 has two unknown parameters that require estimation, $(N \mid G = F, v = j + h)$ and $P(B \mid v = j + h, h)$. Let us first consider the estimation of $(N \mid G = F, v = j + h)$, stated formally in Result 4.4.

Result 4.4: The number of focal group members at valid score category $v = j$ is estimated by the number of focal group members at observed score category $m =$
\(j.\) That is

\[(N | G = F, \nu = j + h) \text{ is estimated by } (N | G = F, m = j + h).\] (4.9)

It is noted that the observed number of focal group examinees at observed score category \(j\) is not an unbiased estimator of the number of focal group examinees at valid score category \(\nu = j\). The extent to which the estimation is biased is contingent upon the target ability distribution of the focal group sample, as well as where the valid score category is in the score distribution. Attempts were made (see Appendix E) to estimate the number of focal group members at valid score category \(\nu = j\) using the information of focal group ability distribution and the placement of the valid score in the score distribution. However, these estimates proved to be highly unstable, and as a result, it was decided to use the observed number of focal group members at \(m\) as the estimate as described in Result 4.4.

We now turn our attention to the estimation of \(P(B | v = j + h, h)\). Once again, \((B | v = j + h, h)\) represents the event that a focal group member at valid score category \(\nu = j + h\) incorrectly responds to \(h\) of the \(k\) biased items due to sub-threshold levels on the nuisance determinant dimension. Let us consider the number of focal group examinees at observed score category \(m = j\) who have been misclassified from each possible higher valid score category. For example, the number of examinees misclassified into observed score category \(j\) from valid score category \(\nu = j + 1\) will be those who missed one and only one item due to bias. Similarly, the number of examinees misclassified into observed score category \(j\) from valid score category \(\nu = j + h\) will be the number of examinees who missed \(h\) and only \(h\) of the \(k\) biased items due to sub-threshold levels of manifest nuisance determinant. The number misclassified can be expected to be distributed as a generalized binomial variable. The
generalized binomial distribution (Stuart & Ord, 1994, Section 5.10) concerns the situation in which an observation is drawn from \( k \) different populations with probabilities of success \( P_1, P_2, \ldots, P_k \), as opposed to the more familiar form of the binomial where \( k \) observations are drawn from the same population with probability of success \( P \). That is, in the case of the binomial distribution all \( k \) trials are sampled from the same population, and thus the probability of success is identical for each trial, while in the generalized binomial each of the \( k \) trials is sampled from a different population, and each population has a unique probability of success. Thus, while in the case of the binomial distribution the probability of \( h = 0, 1, \ldots, k \) successes given \( k \) trials is generated by the terms of the product

\[
(P + Q)^k,
\]

(4.10)

(see Pitman, 1993, p. 80) where \( P \) and \( Q \) denote the probabilities of success and failure on any given trial, the probability of \( h \) successes given \( k \) trials for the generalized binomial distribution is generated by the terms of the product

\[
(P_1 + Q_1)(P_2 + Q_2)\ldots(P_k + Q_k) = \prod_{i=1}^{k} (P_i + Q_i)
\]

(4.11)

where \( i \) refers to any one of the \( k \) trials. Multiplying through all of the terms of the left hand side of Equation 4.11 gives the probability of having \( h \) successes out of \( k \) trials, where \( h = 0, 1, 2, \ldots, k \). An applied example of the generalized binomial distribution is given after Result 4.5. In the context of this research, \( P_i \) is the probability of missing item \( i \) due to a sub-threshold level on the nuisance determinant dimension, which has previously been defined in terms of manifest bias as described in Definition 3.1, and is denoted here by \( \delta_i \). Using this
information, the probability of incorrectly responding to \( h \) of \( k \) items due to sub-threshold levels on the nuisance determinant dimension when \( k \) items are biased can be determined by the generalized binomial distribution. This is stated formally in Result 4.5.

**Result 4.5:** The probability of a focal group examinee at valid score category \( v = j \) incorrectly responding to \( h \) of \( k \) biased items due to sub-threshold levels on the nuisance determinant dimension can be expressed as

\[
P(B|v = j, h) = \sum_{a=1}^{b} \prod_{i=1}^{h} (\delta_i|v = j) \prod_{r=1}^{k-h} (1 - (\delta_r|v = j))
\] (4.12)

where \( k \) is the number of biased items, \( i \) is any one of the \( h \) items incorrectly responded to due to a sub-threshold level of expressed nuisance determinant, \( r \) is any one of the biased items not incorrectly responded to due to a sub-threshold level of expressed nuisance determinant (where \( i \) is not equal to \( r \)), \( a \) is any one of the \( b \) possible combinations of missing \( h \) of \( k \) items due to sub-threshold levels of expressed nuisance determinant, and \((\delta|v = j)\) represents the probability of missing the item due to a sub-threshold level of expressed nuisance determinant for focal group members at valid score category \( j \).

In words, Equation 4.12 computes the probability of being biased against on \( h \) of the \( k \) biased items by computing the probability of being biased against on \( h \) of the \( k \) items for each possible combination of \( h \) biased items. Different items have different associated levels of manifest bias, and thus bias on different combinations of the \( h \) items will have different probabilities of occurrence. Thus, it is necessary to consider each combination individually. Once the probability of missing each combination of \( h \) biased items due to sub-threshold
levels on the nuisance determinant dimension has been determined, the summation used in Equation 4.12 is over the individual probability for each combination to arrive at a total probability of missing $h$ items due to sub-threshold levels of manifest nuisance determinant.

The relationship of Equation 4.12 to Equation 4.11 may not be immediately apparent, but can be described as follows. Consider the case of expanding the terms of Equation 4.11 for the case of having 3 biased items with probability of being biased against on item $i$ equal to $\delta_i$ and probability of not being biased against on item $i$ equal to $(1 - \delta_i)$. Then, using Equation 4.11 we have

$$\prod_{i=1}^{k} \delta_i (1 - \delta_i) = \left[ \delta_1 + (1 - \delta_1) \chi_2 + (1 - \delta_2) \chi_3 + (1 - \delta_3) \right].$$

Expanding the right hand side we have the following terms, which have been grouped according to the number of items being biased against ($h$).

For $h = 3$: $\delta_1 \delta_2 \delta_3$

For $h = 2$: $\delta_1 \delta_2 (1 - \delta_3) + \delta_1 \delta_3 (1 - \delta_2) + \delta_2 \delta_3 (1 - \delta_1)$

For $h = 1$: $\delta_1 (1 - \delta_2) (1 - \delta_3) + \delta_2 (1 - \delta_1) (1 - \delta_3) + \delta_3 (1 - \delta_1) (1 - \delta_2)$

For $h = 0$: $(1 - \delta_1) (1 - \delta_2) (1 - \delta_3)$

The top row provides the probability of being biased against on each of the three biased items, the second row provides the probability of being biased against on any two of the three biased items, etc. Equation 4.12 computes any one of these rows for focal group members with valid test score $v = j + h$.

As a final step in the chain of estimation procedures, we require (for Equation 4.12) an estimate of the manifest bias of the item at valid score category $v = j$. This estimate is
given in Result 4.6.

**Result 4.6:** An estimate of the manifest bias at valid score category $v = j$ is given by

$$
(\hat{\delta} | v = j) = (\bar{Y} | G = R, m = j) - (\bar{Y} | G = F, m = j).
$$

(4.13)

**Proof:** From Definition 3.1 it can be stated that

$$
(\delta | v = j) = P(\Theta = 1 | G = F, v = j) - P(Y = 1|G = F, v = j).
$$

(4.14)

Using the Result 3.2, and the assumption that for reference population members $m = v$ (see Equations 4.2 and 4.3), we have

$$
(\delta | v = j) = P(Y = 1 | G = R, m = j) - P(Y = 1 | G = F, v = j).
$$

(4.15)

The value $P(Y = 1 | G = F, v = j)$ is unknown, but is estimated here using $(\bar{Y} | G = F, m = j)$.

Note that because the probability of correct response for the reference group at observed score category $m = j$ is used to estimate the focal group probability of success on the target ability dimension at valid score category $v = j$ (Equation 4.15), the estimate of manifest bias in Equation 4.13 is dependent on the equality of group ability distributions.

Using the findings of Results 4.1 to 4.6, it is now possible to estimate the number of successes on the target ability dimension for the focal group members at observed score category $j$, as outlined in Equation 4.1. The sequence of steps used to conduct this estimation is summarized in Table 4.1. Note that, as shown in Table 4.1, in practice the estimation procedures flow from Result 4.6 to Result 4.1.
Table 4.1
Steps Used in Estimating The Valid Focal Group Item Difficulty at Observed Score $m = j$

<table>
<thead>
<tr>
<th>Step #</th>
<th>Result #</th>
<th>Description of the Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.6</td>
<td>Estimate the manifest bias at each valid score category using &lt;br&gt;$$\hat{\delta}_j = (\bar{Y}</td>
</tr>
<tr>
<td>2</td>
<td>4.5</td>
<td>Using the estimate of manifest bias given in Step #1, estimate the probability of being biased against on $h$ of the $k$ biased items using &lt;br&gt;$$P(B</td>
</tr>
<tr>
<td>3</td>
<td>4.4</td>
<td>Estimate $(N</td>
</tr>
<tr>
<td>4</td>
<td>4.3</td>
<td>Using information obtained in Steps #2 and #3, estimate $E(N</td>
</tr>
<tr>
<td>5</td>
<td>4.2</td>
<td>Estimate $E(\Theta</td>
</tr>
<tr>
<td>6</td>
<td>4.1</td>
<td>Using information obtained in Step #5, estimate &lt;br&gt;$$E \left[ \sum_{p=1}^{N} (\Theta</td>
</tr>
<tr>
<td>7</td>
<td>Equation 4.1</td>
<td>Using the information obtained from Step #6, estimate the probability of focal group success on the target ability dimension using &lt;br&gt;$$P(\Theta = {G = F, m = j}) = \frac{\sum_{p=1}^{N_{fj}} (\Theta_p</td>
</tr>
</tbody>
</table>
Chapter Five

Assessing the Performance of the MH Adjustment

Introduction

Chapters Three and Four contain the development of theory and estimation procedures for an adjustment to the MH procedure, which is intended to correct for the effects of CIM. This chapter is the report of an experimental assessment of the adjustment by means of a simulation study.

Method

The simulations were conducted by creating an artificial test of 40 dichotomous items. This choice of length was predicated upon two factors. First, unlike the simulation study conducted in Chapter Two where it was of interest to determine how the effects of CIM might be dependent on the test length, the effect of test length on the performance of the MH adjustment was judged to be of little interest; it is assumed that the performance of the MH adjustment is largely independent of the number of items on the test. As a result, it was decided that only one test length need be considered to gain a sense of how well the adjustment performs. Second, the computer memory resources required to run simulations of lengthy tests exceeded that available in the software used to conduct the simulations (SAS). This limitation resides primarily in the limit of array elements of approximately 36,000 permitted by SAS. Due to the computational complexity of the adjustment, simulations using tests of 60 items, as used in Chapter Two, exceeded this number. Thus, a test length was
sought which would be representative of typical testing situations, but small enough to permit the simulations given the software resource limitations. For these reasons, a 40-item test was selected. The use of a 40-item test is supported by previous simulation studies of DIF (Clauser, Mazor & Hambleton, 1993; Swaminathan & Rogers, 1990).

The procedures used to generate simulated responses were identical to those described in Chapter Two. The five factors examined in this study were: (1) number of non-studied items containing DIF, (2) magnitude of DIF in the non-studied items, (3) magnitude of DIF in the studied item, (4) number of examinees in each group, and (5) equality of the means of the focal and reference group ability distributions. Each of these factors is discussed below.

**Factor 1: Number of Contaminating Items.** Approximately 5%, 10%, and 20% of the non-studied items contained DIF. These percentages correspond to 2, 4, and 8 non-studied items containing DIF.

**Factor 2: Magnitude of DIF in Contaminating Items.** Two levels of DIF in the non-studied items were considered; $t = 0.3$ and 0.6. As described in Chapter Two, these values correspond to moderate and high levels of DIF.

**Factor 3: Magnitude of DIF in the Studied Item.** Two levels of DIF were considered in the studied item; $t = 0.0$ and 0.6. These levels were identical to those used in Chapter Two, and permit the assessment of the Type I error rate and power of the adjustment.

**Factor 4: Number of Examinees.** Two levels of group size were considered; 250 and 1000. The condition of 250 examinees is critical to examining the stability of the adjustment when only small samples are available.

**Factor 5: Equality of Group Ability Means.** Consideration was given to two levels of
difference in the mean of the focal and reference group ability distributions. The first level was a zero difference between the means of the group ability distributions ($\mu_R = \mu_F = 0.0$). The second level places the mean focal group ability distribution one standard deviation below that of the reference group ($\mu_R = 0.0$, $\mu_F = -1.0$).

All conditions in which non-studied items contain DIF were crossed, giving a total of 48 conditions. For each condition, 1000 replications were run. For all conditions, the significance of DIF was assessed at a 0.05 level of significance. Four additional conditions (for each combination of group size and equality of group ability distribution means) were run in which no non-studied items contained DIF. The efficiency of the adjustment was assessed by comparing the mean adjusted $MHD$-$DIF$, the power and Type I error rate of the adjusted $MH\chi^2$, and the adjusted proportion of Category C items to that of the control condition in which none of the non-studied items contained DIF. In addition, since the z-score for the $MHD$-$DIF$ value is often used to assess DIF (as in the ETS classification system), the ratio of the adjusted $MHD$-$DIF$ value to its standard deviation was also used to assess the performance of the adjustment.

One final note concerning the form of the adjustment used in this paper. While this adjustment can be used to correct for the effects of all items containing DIF (studied and non-studied) the form of the adjustment presented here corrects only for the contamination arising from DIF in the non-studied items. The decision to correct for only the non-studied items was based on two factors, previously discussed in the Method section of Chapter Two. First, contamination caused by the studied item will affect all studied items having moderate to high levels of DIF relatively equally. As a consequence, the contamination incurred by the
studied item’s DIF will not substantially affect the relative rating of DIF in items. Second, the purpose of the simulation study was to assess the performance of the adjustment. To this end, the distribution of adjusted MH values was compared to a control condition in which no contamination existed. Unfortunately, there is no empirical way of obtaining a “true” control condition in which the studied item contains DIF without CIM existing, making it impossible to control for the CIM caused by the studied item. Thus, the performance of the adjustment was assessed according to how well it controlled for CIM from only non-studied items. If it was found to perform well in this situation, then it can be inferred that it would perform satisfactorily for the studied item as well.

Results

Table 5.1 displays the results for the condition in which sample size equaled 1000 and the reference and focal group ability distributions had equal means. The top line represents the values obtained for the control condition in which no DIF was introduced into the non-studied items. The results indicate that the mean adjusted MHD-DIF values remained near the control value under all levels of contamination, relative to the mean observed MHD-DIF values. The deviation of the mean of the adjusted MHD-DIF from the control value increased as contamination increased; the absolute value of the deviation equaling 0.047 under the smallest levels of contamination, and increasing to 0.068 under the largest levels of contamination. Note, however, that the mean adjusted MHD-DIF value of -1.427 was a large improvement over the mean observed MHD-DIF value of -1.192. This result is displayed graphically in Figure 5.1 for the case in which t = 0.6 for contaminating non-studied items.
While the mean adjusted *MHD-DIF* value under the highest level of contamination was lower than the corresponding control value, it is interesting to note that the mean adjusted z-scores remained near the control value across all levels of contamination. In contrast, the mean observed z-scores decreased as contamination increased.

Table 5.1

<table>
<thead>
<tr>
<th>%</th>
<th>t</th>
<th>MHD-DIF</th>
<th>z-score</th>
<th>Power</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>-1.495</td>
<td>5.567</td>
<td>0.97</td>
<td>0.19</td>
<td>0.32</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
<td>-1.419</td>
<td>5.291</td>
<td>0.96</td>
<td>0.21</td>
<td>0.40</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.448</td>
<td>5.435</td>
<td>0.96</td>
<td>0.20</td>
<td>0.37</td>
<td>0.44</td>
</tr>
<tr>
<td>4</td>
<td>0.3</td>
<td>-1.408</td>
<td>5.240</td>
<td>0.96</td>
<td>0.22</td>
<td>0.37</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.468</td>
<td>5.549</td>
<td>0.96</td>
<td>0.20</td>
<td>0.35</td>
<td>0.46</td>
</tr>
<tr>
<td>8</td>
<td>0.3</td>
<td>-1.338</td>
<td>4.904</td>
<td>0.94</td>
<td>0.24</td>
<td>0.41</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.446</td>
<td>5.498</td>
<td>0.95</td>
<td>0.21</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>2</td>
<td>0.6</td>
<td>-1.393</td>
<td>5.222</td>
<td>0.94</td>
<td>0.23</td>
<td>0.35</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.458</td>
<td>5.500</td>
<td>0.95</td>
<td>0.19</td>
<td>0.33</td>
<td>0.48</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
<td>-1.341</td>
<td>4.999</td>
<td>0.95</td>
<td>0.25</td>
<td>0.39</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.472</td>
<td>5.650</td>
<td>0.96</td>
<td>0.19</td>
<td>0.34</td>
<td>0.47</td>
</tr>
<tr>
<td>8</td>
<td>0.6</td>
<td>-1.192</td>
<td>4.499</td>
<td>0.93</td>
<td>0.32</td>
<td>0.49</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.427</td>
<td>5.607</td>
<td>0.95</td>
<td>0.20</td>
<td>0.35</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Note. % refers to the number of non-studied items containing DIF, *t* represents the magnitude of DIF introduced into the non-studied items, *MHD-DIF* refers to the mean value of *MHD-DIF*, Power refers to the proportion of trials having a significant value of $MH\chi^2$, and A, B and C represent the proportion of items classified as Category A, B, and C respectively. The bold type indicates the adjusted values, and regular type the uncorrected values.
Considering still Table 5.1, we note that for all levels of contamination there was a decrease of between 0.02 and 0.05 in the proportion of items flagged as Category C using the adjusted MHD-DIF values relative to the control levels. These results are a large improvement over the observed MHD-DIF results. The adjustment consistently accounted for approximately 80% of the observed MHD-DIF underestimation in the proportion of items flagged as Category C. Furthermore, under the largest level of contamination the proportion of items flagged as Category C by the adjustment was more than double that of the observed MHD-DIF. These results are presented graphically in Figure 5.1 for the condition in which $t = 0.6$ for contaminated non-studied items.

Table 5.2 displays the results for the conditions in which group sample size equaled 250 and the reference and focal group ability distributions had equal means. The results indicate that while there is the same general pattern as was found for the group sizes of 1000, the adjustment becomes less effective for group sizes of 250. Specifically, as contamination increases, the adjustment increasingly underestimates the control MHD-DIF value. This underestimation grows to approximately 0.10 under high levels of contamination. The underestimation of the control MHD-DIF values by the adjusted MHD-DIF values was accompanied by a parallel decrease in power of the $MHD^2$, the power decreasing to 0.65 from a control condition value of 0.74. Although there was a decrease in power of the adjusted MHD-DIF values, this power was still greater than that of the observed MHD-DIF values. Note that the proportion of items classified as Category C for the adjusted MHD-DIF exceeded that of the control condition under the highest level of contamination.
Figure 5.1

A Comparison of the Control, Observed and Adjusted Results

Figure 5.1. The mean control, adjusted and observed $MHD-DIF$ values (absolute values) and proportion of items classified as Category C are displayed as a function of percentage of non-studied items containing DIF ($t = 0.6$). Standard errors are on the order of 0.008 for the mean $MHD-DIF$ values, and 0.01 for the proportion of items classified as Category C. The results displayed are for conditions in which sample size equals 1000, and group ability distribution means were equal. In all conditions, the studied item had an induced DIF of $t = 0.6$. 
Table 5.2
Mean MHD-DIF Values, and Detection Rates for Sample Sizes of 250, and Group Ability Distributions Having Equal Means

<table>
<thead>
<tr>
<th>%</th>
<th>t</th>
<th>MHD-DIF</th>
<th>z-score</th>
<th>Power</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>-1.529</td>
<td>2.756</td>
<td>0.74</td>
<td>0.25</td>
<td>0.57</td>
<td>0.18</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
<td>-1.499</td>
<td>2.664</td>
<td>0.70</td>
<td>0.29</td>
<td>0.55</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.486</td>
<td>2.668</td>
<td>0.70</td>
<td>0.31</td>
<td>0.52</td>
<td>0.18</td>
</tr>
<tr>
<td>4</td>
<td>0.3</td>
<td>-1.425</td>
<td>2.567</td>
<td>0.68</td>
<td>0.32</td>
<td>0.55</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.443</td>
<td>2.662</td>
<td>0.68</td>
<td>0.33</td>
<td>0.49</td>
<td>0.18</td>
</tr>
<tr>
<td>8</td>
<td>0.3</td>
<td>-1.366</td>
<td>2.491</td>
<td>0.65</td>
<td>0.33</td>
<td>0.55</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.417</td>
<td>2.800</td>
<td>0.65</td>
<td>0.35</td>
<td>0.47</td>
<td>0.19</td>
</tr>
<tr>
<td>2</td>
<td>0.6</td>
<td>-1.429</td>
<td>2.558</td>
<td>0.67</td>
<td>0.33</td>
<td>0.53</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.446</td>
<td>2.631</td>
<td>0.67</td>
<td>0.33</td>
<td>0.51</td>
<td>0.17</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
<td>-1.341</td>
<td>2.430</td>
<td>0.63</td>
<td>0.35</td>
<td>0.53</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.438</td>
<td>2.738</td>
<td>0.67</td>
<td>0.34</td>
<td>0.47</td>
<td>0.19</td>
</tr>
<tr>
<td>8</td>
<td>0.6</td>
<td>-1.201</td>
<td>2.187</td>
<td>0.54</td>
<td>0.44</td>
<td>0.48</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.446</td>
<td>2.820</td>
<td>0.65</td>
<td>0.33</td>
<td>0.46</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Note. See Table 5.1 for a description of the headers.

Tables 5.3 and 5.4 display the results for the conditions in which the group ability distributions were unequal, and group sizes equaled 1000 and 250, respectively. These results parallel those of the conditions in which group ability distribution means were equal.

The adjustment performed better with large sample sizes than with small. In particular, when sample sizes were large, the power of the adjusted $MH\chi^2$ decreased slightly, as did the proportion of items identified as Category C. When sample sizes were small, the decrease in power was more pronounced, reaching 0.10 under the most severe level of contamination.

While the proportion of items flagged as Category C in the small sample case was unaffected
by increasing contamination, the proportion of items flagged as Category B decreased from 0.44 to 0.36.

Table 5.3
Mean MHD-DIF Values, and Detection Rates for Sample Sizes of 1000, and Group Ability Distributions Having Unequal Means

<table>
<thead>
<tr>
<th>%</th>
<th>t</th>
<th>MHD-DIF</th>
<th>z-score</th>
<th>Power</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>-1.247</td>
<td>4.537</td>
<td>0.85</td>
<td>0.37</td>
<td>0.28</td>
<td>0.35</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>-1.185</td>
<td>4.272</td>
<td>0.81</td>
<td>0.41</td>
<td>0.28</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.210</td>
<td>4.387</td>
<td>0.82</td>
<td>0.40</td>
<td>0.28</td>
<td>0.33</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
<td>-1.102</td>
<td>4.200</td>
<td>0.80</td>
<td>0.40</td>
<td>0.31</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.207</td>
<td>4.408</td>
<td>0.82</td>
<td>0.38</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>20</td>
<td>0.3</td>
<td>-1.088</td>
<td>3.955</td>
<td>0.78</td>
<td>0.43</td>
<td>0.33</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.173</td>
<td>4.346</td>
<td>0.80</td>
<td>0.39</td>
<td>0.30</td>
<td>0.31</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>-1.144</td>
<td>4.142</td>
<td>0.79</td>
<td>0.41</td>
<td>0.31</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.195</td>
<td>4.372</td>
<td>0.81</td>
<td>0.39</td>
<td>0.30</td>
<td>0.31</td>
</tr>
<tr>
<td>10</td>
<td>0.6</td>
<td>-1.072</td>
<td>3.896</td>
<td>0.76</td>
<td>0.47</td>
<td>0.28</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.172</td>
<td>4.346</td>
<td>0.79</td>
<td>0.42</td>
<td>0.26</td>
<td>0.32</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>-.981</td>
<td>3.558</td>
<td>0.74</td>
<td>0.50</td>
<td>0.33</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.153</td>
<td>4.328</td>
<td>0.79</td>
<td>0.41</td>
<td>0.29</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Note. See Table 5.1 for a description of the headers.
Table 5.4
Mean MHD-DIF Values, and Detection Rates for Sample Sizes of 250, and Group Ability Distributions Having Unequal Means

<table>
<thead>
<tr>
<th>Cont</th>
<th>t</th>
<th>MHD-DIF</th>
<th>z-score</th>
<th>Power</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>-1.258</td>
<td>2.248</td>
<td>0.57</td>
<td>0.42</td>
<td>0.44</td>
<td>0.14</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>-1.224</td>
<td>2.212</td>
<td>0.54</td>
<td>0.44</td>
<td>0.43</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.250</td>
<td>2.312</td>
<td>0.54</td>
<td>0.44</td>
<td>0.40</td>
<td>0.15</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
<td>-1.165</td>
<td>2.105</td>
<td>0.50</td>
<td>0.48</td>
<td>0.41</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.209</td>
<td>2.278</td>
<td>0.52</td>
<td>0.47</td>
<td>0.37</td>
<td>0.16</td>
</tr>
<tr>
<td>20</td>
<td>0.3</td>
<td>-1.089</td>
<td>2.011</td>
<td>0.47</td>
<td>0.52</td>
<td>0.38</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.167</td>
<td>2.400</td>
<td>0.51</td>
<td>0.48</td>
<td>0.36</td>
<td>0.17</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>-1.136</td>
<td>2.044</td>
<td>0.49</td>
<td>0.50</td>
<td>0.40</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.190</td>
<td>2.199</td>
<td>0.52</td>
<td>0.46</td>
<td>0.41</td>
<td>0.13</td>
</tr>
<tr>
<td>10</td>
<td>0.6</td>
<td>-1.118</td>
<td>2.029</td>
<td>0.47</td>
<td>0.51</td>
<td>0.40</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.209</td>
<td>2.299</td>
<td>0.52</td>
<td>0.47</td>
<td>0.37</td>
<td>0.16</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>-0.931</td>
<td>1.760</td>
<td>0.38</td>
<td>0.60</td>
<td>0.36</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.133</td>
<td>2.311</td>
<td>0.47</td>
<td>0.50</td>
<td>0.36</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Note. See Table 5.1 for a description of the headers.

Table 5.5 displays the results for the null conditions in which no DIF was introduced into the studied item when group size equaled 1000. As with the previous results of this chapter, the first row of the table represents a control condition in which no contamination existed. When group ability distribution means were equal, the mean value of the observed MHD-DIF increased from the control value of -0.016 to over 0.30 as contamination increased. In contrast, the mean of the adjusted MHD-DIF remained near the control level, increasing to only 0.045 under the condition of maximal contamination. The adjusted $MH\chi^2$ also maintained Type I error rates near the nominal level of 0.05, which was a great
improvement over those observed for the unadjusted $MH\chi^2$ which displayed Type I error rates as high as 0.19 under the condition of maximal contamination. Similar results were obtained when group ability distributions had unequal means.

Table 5.5
Simulation Results for the Conditions in which no DIF is Introduced in the Studied Item, and Group Size is 1000

<table>
<thead>
<tr>
<th>%</th>
<th>t</th>
<th>Unequal Means</th>
<th></th>
<th>Equal Means</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$MHD$-$DIF$</td>
<td>Type I</td>
<td>$MHD$-$DIF$</td>
<td>Type I</td>
</tr>
<tr>
<td>0</td>
<td>0.0</td>
<td>-0.016</td>
<td>0.08</td>
<td>-0.007</td>
<td>0.05</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>0.026</td>
<td>0.07</td>
<td>0.041</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.001</td>
<td>0.07</td>
<td>0.008</td>
<td>0.05</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>0.046</td>
<td>0.07</td>
<td>0.122</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.008</td>
<td>0.08</td>
<td>0.014</td>
<td>0.10</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
<td>0.062</td>
<td>0.08</td>
<td>0.078</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.006</td>
<td>0.08</td>
<td>0.013</td>
<td>0.05</td>
</tr>
<tr>
<td>10</td>
<td>0.6</td>
<td>0.100</td>
<td>0.09</td>
<td>0.156</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.004</td>
<td>0.07</td>
<td>0.012</td>
<td>0.04</td>
</tr>
<tr>
<td>20</td>
<td>0.3</td>
<td>0.129</td>
<td>0.09</td>
<td>0.176</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.041</td>
<td>0.08</td>
<td>0.056</td>
<td>0.08</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>0.237</td>
<td>0.15</td>
<td>0.301</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.062</td>
<td>0.09</td>
<td>0.045</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Note. % refers to the number of non-studied items containing DIF, $t$ represents the magnitude of DIF introduced into the non-studied items, $MHD$-$DIF$ represents the mean of the absolute value of the $MHD$-$DIF$ index, and Type I represents the Type I error rate of $MH\chi^2$. The bold type indicates the adjusted values, and regular type the uncorrected values.
Table 5.6 displays the results for the conditions in which no DIF was introduced into the studied item when group size equaled 250. For conditions in which the group ability distributions had equal means, the adjustment maintained mean $MHD$-$DIF$ values near the control value. However, the Type I error rate of the adjusted $MH^2$ rose to 0.09 as contamination increased. In addition, the Type I error rate of the adjusted $MH^2$ exceeded that of the unadjusted $MH^2$, at times by more than 0.06. A similar finding existed for the conditions in which group ability distributions had unequal means.

<table>
<thead>
<tr>
<th>%</th>
<th>$t$</th>
<th>Unequal Means $MHD$-$DIF$</th>
<th>Type I</th>
<th>Equal Means $MHD$-$DIF$</th>
<th>Type I</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>-0.025</td>
<td>0.04</td>
<td>-0.045</td>
<td>0.04</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>0.007</td>
<td>0.04</td>
<td>0.052</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.003</td>
<td>0.06</td>
<td>0.060</td>
<td>0.06</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>0.008</td>
<td>0.04</td>
<td>0.078</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.031</td>
<td>0.05</td>
<td>0.026</td>
<td>0.06</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
<td>0.038</td>
<td>0.03</td>
<td>0.078</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.003</td>
<td>0.08</td>
<td>0.052</td>
<td>0.08</td>
</tr>
<tr>
<td>10</td>
<td>0.6</td>
<td>0.132</td>
<td>0.05</td>
<td>0.151</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.028</td>
<td>0.08</td>
<td>0.039</td>
<td>0.07</td>
</tr>
<tr>
<td>20</td>
<td>0.3</td>
<td>0.117</td>
<td>0.05</td>
<td>0.160</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.046</td>
<td>0.13</td>
<td>0.048</td>
<td>0.09</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>0.230</td>
<td>0.05</td>
<td>0.308</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.046</td>
<td>0.11</td>
<td>0.047</td>
<td>0.07</td>
</tr>
</tbody>
</table>

*Note.* See Table 5.5 for a description of headers.
Discussion

The results of the simulation study indicate that the performance of the adjustment depends on: (1) the magnitude of contamination, (2) the sample size, and (3) the equality of the reference and focal group ability distribution means. The performance of the adjustment in accounting for the bias in the testing of DIF by the observed $MHD-DIF$ value and $MH\chi^2$ was greatest when sample sizes were large ($N = 1000$) and group ability distribution means were equal. Performance was degraded when sample sizes were small ($N = 250$) or group ability distribution means were unequal.

The use of the adjustment with small sample sizes is cautioned due to two findings. First, when sample sizes were small, the proportion of items flagged as Category C often exceeded the control level. Second, when sample sizes were small, the adjustment had Type I error rates that were higher than the unadjusted values. These findings suggest that the estimation procedures are particularly unstable with small sample sizes.

In interpreting the results, several limitations should be considered which could degrade the validity and performance of the adjustment in applied testing situations. The first limitation concerns determining which items to adjust for. Within the context of this simulation study, the set of items to be adjusted for is established a priori. However, in real test development situations, this luxury does not exist, and a criterion would need to be developed to determine which items are to be considered in the adjustment. For example, this criterion could be based on the ETS classification system, whereby all items in Categories B and C are corrected for. Future investigations into the feasibility of the adjustment should study the impact of the adjustment when such a criterion is used.
A related limitation concerns the relationship between bias and DIF. As described in Chapter One, DIF is a necessary, but not sufficient, condition for bias. The implication of this is that in order to obtain a “contamination-free” value of $MHD-DIF$, one must adjust only for those items containing bias. If some of the items containing DIF were in fact not biased, then the adjustment would overcorrect the MH value.

A third limitation of the results is that the performance of the adjustment has been assessed using simulated data. While the use of simulated data has the advantage of allowing the adjustment to be compared to a control value (in which no contamination from non-studied items exists), the use of simulated data means that the results are not necessarily generalizable to real testing data, particularly data which have a poor fit to a three parameter logistic regression IRT model.

A final limitation concerns the assumptions made in the statistical formulation of the adjustment presented in Chapters Three and Four. The assumptions most likely to be violated in practice are the equality of target ability distributions of the reference and focal populations (stated in Chapter Four), and the independence of the success on the target ability and nuisance determinant dimensions (stated in Chapter Three). Although the precise effects of violating these assumptions on the performance of the MH adjustment are unknown, the extent to which these assumptions are met may play a role in determining the applicability of the adjustment.
Chapter Six
Developing MB-DIF

Introduction

In Chapter Four, an adjustment to the observed MH value was proposed to correct for the effects of CIM. Unfortunately, this adjustment suffers from two limitations. First, under the most severe levels of contamination, it was found that the adjustment corrects for only about 80% of the underestimation attributable to CIM. Second, the adjustment is computationally intensive, requiring first the calculation of the uncorrected MH values to determine which items to adjust for, then the corrected reference group statistics (\( \tilde{A} \) and \( \tilde{B} \)), followed by the calculation of the adjusted MH values. These limitations present obstacles to the practical utilization of the adjustment that may limit the effectiveness of the adjustment as a solution to the problem of CIM. In this chapter I explore a second possible solution to the problem of CIM; namely, the development of a DIF detection procedure that is robust to the effects of CIM, while maintaining power and Type I error rates comparable to the MH procedure when no CIM exists. The theoretical development of such an alternative DIF detection procedure, denoted by MB-DIF, is the topic of this chapter.

Defining MB-DIF

An alternative DIF detection method was sought that is robust to the effects of CIM, but maintains power and Type I error rates comparable to those of the MH\( \chi^2 \). I postulated
that such a robust measure could be obtained from an estimate of the manifest bias of the item. To review a portion of the theory developed in Chapter Three, manifest bias for any fixed value of \( \theta \) and \( \eta \) is defined as

\[
\delta(\theta, \eta) = P(\Theta = 1|\theta) - P(Y = 1|\theta, \eta) .
\]

(6.1)

The manifest bias for the focal population members at observed score category \( m = j \) can be obtained by

\[
\delta_j = P(\Theta = 1|G = F, m = j) - P(Y = 1|G = F, m = j) .
\]

(6.2a)

The manifest bias in Equation 6.2a can be estimated using

\[
\hat{\delta}_j = \hat{P}(\Theta = 1|G = F, m = j) - (\hat{Y}|G = F, m = j) .
\]

(6.2b)

Procedures for estimating the first term on the right hand side of Equation 6.2b were developed in Chapter Four.

For the calculation of MB-DIF I seek a value of manifest bias that is representative of the bias experienced by all focal group members. Since the manifest bias expressed in Equation 6.2a cannot be expected to be constant across all observed score categories, it is necessary to obtain an estimate of the manifest bias over all \( n \) observed score categories. An index of the manifest bias in the entire focal group can be obtained by creating a weighted composite of manifest bias across all \( n \) observed score categories, given by

\[
\delta = \sum_{j=1}^{n} w_j \delta_j .
\]

(6.3)

where \( j \) is a particular value of the observed score category \( m \), and \( w_j \) is the weight assigned to score category \( j \). To keep the weighted composite on the same scale as the manifest bias
observed at any score category (bounded between 0 and 1) the weight assigned to each score category can be adjusted by dividing by the sum of the weights across all score categories.

This can be expressed by

$$\delta = \sum_{j=1}^{n} \delta_j \left( \frac{w_j}{\sum_{j=1}^{n} w_j} \right)$$  \hspace{1cm} (6.4a)$$

$$\sum_{j=1}^{n} w_j \delta_j$$

$$= \frac{\sum_{j=1}^{n} w_j \delta_j}{\sum_{j=1}^{n} w_j}$$  \hspace{1cm} (6.4b)$$

where $n$ is the number of score categories, and $w_j$ is the weight associated with the manifest bias of each score category. It was decided to weight the manifest bias of each score category by a measure of the between-group information contained in the associated score category.

As described for the $\alpha_{MH}$ in Chapter One, a measure of the between-group information at observed score category $j$ ($I_j$) can be obtained by

$$I_j = \frac{N_{Rj} N_{Fj}}{N_{Rj} + N_{Fj}}$$  \hspace{1cm} (6.5)$$

and the proportion of the total between-group information at observed score category $j$ ($w_j$) is given by
The weight shown in Equation 6.6 is equivalent to the proportion of total sums of squares of the discrete variable $G$ at observed score category $j$, where $G$ codes group membership for the reference and focal groups. This is shown in Result 6.1.

**Result 6.1:** The weight assigned to each score category using

$$I_j = \frac{N_{Rj} \cdot N_{Fj}}{N_{Rj} + N_{Fj}}$$

is equivalent to the total sum of squared deviations of the discrete variable $G$ which codes for group membership for the reference and focal groups.

**Proof:** Let $G$ be the dichotomous variable that assigns the value of 1 for the reference group and 0 for the focal group. Then the weight shown in Equation 6.5 for any sample of focal and reference group members can be expressed as

$$\frac{N_R N_F}{N_R + N_F} = \left( \frac{N_R}{N_R + N_F} \right) \left( \frac{N_F}{N_R + N_F} \right) (N_R + N_F)$$

$$= P(G = F) P(G = R) (N_R + N_F)$$

$$= [1 - P(G = R)] P(G = R) (N_R + N_F)$$

$$= \left[ P(G = R) - \left( P(G = R) \right)^2 \right] (N_R + N_F)$$
\[
\begin{align*}
&= \left[ E(G^2) - (E(G))^2 \right] (N_R + N_F) \\
&= \sigma_G^2 (N_R + N_F) \\
&= \frac{\sum_{i=1}^{N_R+N_F} [G_i - E(G)]^2}{N_R + N_F} (N_R + N_F), \\
&= \sum_{i=1}^{N_R+N_F} [G_i - E(G)]^2.
\end{align*}
\]

That is, the weight shown in Equation 6.6 for any one score category is equivalent to the proportion of the total sum of squared deviations of the group variable \( G \) at that score category, which can be interpreted as the proportion of the total between-group information contained at that score category (Mantel & Haenszel, p. 732).

The above weighting procedure has several advantages over the traditional weighting procedure of using the proportion of examinees (either focal group, or combined focal and reference group) at each score category. First, it weights each score category according to the total number of examinees in the category; for a given level of spread of focal and reference group proportions within the category, this weighting procedure will give more weight to a category having more total examinees. For example, more importance will be assigned to a category having 50 reference and 60 focal group members than one having 25 reference and 30 focal group members. Second, this weighting procedure weights each score category according to the spread of reference and focal group members within the category; more
weight is assigned to score categories having equal proportions of reference and focal group members. For example, a score category with 10 focal group members and 90 reference group members will be given less weight than one having 50 members from each group, even though in both cases there are a total of 100 members. As a consequence, the weighting scheme shown in Equation 6.6 will likely result in more desirable levels of power and Type I errors than would be found using the traditional weighting schemes such as the number of focal group members at that score category, or the combined number of focal and reference group members at that score category.

An estimate of the manifest bias shown in Equation 6.4b is given by

\[ \hat{\delta} = \frac{\sum_{j=1}^{n} w_j \hat{\delta}_j}{\sum_{j=1}^{n} w_j} \]  

(6.8)

where \( \hat{\delta}_j \) is given by Equation 6.2b. This estimate of the manifest bias across all focal group members is the MB-DIF\(^1\) statistic. That is

\(^1\) Note that in this form MB-DIF is nearly identical to STNDP-DIF of Dorans and Kulick (1986), with the exception that MB-DIF corrects the reference group probability of success at \( m \) for the effects of CIM. While MB-DIF is a function of the difference in the item-test regressions of the observed and valid focal group probabilities of success, STNDP-DIF is a function of the difference in the item-test regressions of the observed focal and reference group probabilities of success. This relationship was only recently discovered. Originally, manifest bias was derived in a fashion that held no obvious resemblance to STNDP-DIF. It was later discovered that manifest bias could be represented in a more parsimonious fashion by Equation 3.5. As a consequence of this direct relationship, it is apparent that MB-DIF is simply an adjusted form of STNDP-DIF, the adjustment correcting for the CIM-induced difference in the ability distributions of focal and reference group members having the same total test score. Appendix D presents a formal proof of this relationship.
An estimate of the variance for MB-DIF can be expressed as

\[
MB - DIF = \frac{\sum_{j=1}^{n} w_j \hat{\delta}_j}{\sum_{j=1}^{n} w_j}.
\]  

(6.9)

The standard error can be obtained from the square root of Equation 6.10b. With the exception of the weight given to each score category, this standard error is comparable to that proposed by Dorans and Holland (1993, p. 50) for STNDP-DIF.
Chapter Seven

Assessing the Performance of MB-DIF

Introduction

The theory and estimation procedures of MB-DIF were developed in Chapter Six. The performance of MB-DIF was assessed empirically through the use of a simulation study. The results of this empirical investigation is the topic of this chapter.

Method

The simulations were conducted by creating an artificial test of 40 dichotomous items. A rationale for using a 40-item test was presented in the Method section of Chapter Five. The procedures used to generate simulated responses are identical to those described in Chapter Two. The five factors examined in this study were: 1) number of non-studied items containing DIF, 2) magnitude of DIF in the studied item, 3) magnitude of DIF in the non-studied items, 4) number of examinees in each group, and 5) the means of the focal and reference group ability distributions. Each of these factors is discussed below.

Factor 1: Number of Contaminating Items. Approximately 5%, 10%, and 20% of the non-studied items contained DIF. These percentages correspond to 2, 4, and 8 non-studied items, respectively.

Factor 2: Magnitude of DIF in the Studied Item. Two levels of DIF were considered in the studied item, $t = 0.0$ and 0.6. These levels of DIF allowed the investigation of Type I error rate and power, respectively.
Factor 3: Magnitude of DIF in Contaminating Items. Two levels of DIF in the non-studied items were considered; $t = 0.3$, and 0.6. As described in Chapter 2, these values correspond to moderate and high levels of DIF.

Factor 4: Number of Examinees. Two group sizes were considered: 250 and 1000 for each of the reference and focal groups. The condition of 250 examinees is critical to examining the performance of MB-DIF under the condition of a small sample.

Factor 5: Equality of Group Ability Distributions. Two levels of difference in the means of the ability distributions for the focal and reference groups were studied. The first level specified a difference of zero between the group means ($\mu_R = \mu_F = 0.0$), and the second level placed the mean focal group ability one standard deviation below that of the reference group ($\mu_R = 0.0$, $\mu_F = -1.0$).

All conditions in which non-studied items contain DIF were crossed, giving a total of 48 conditions. For each condition, 1000 replications were run. An additional set of four conditions (one for each of the four possible combinations of test length and equality of group ability distribution means) in which $t = 0.0$ for all non-studied items was run as control conditions in which no contamination due to non-studied items was present. The performance of MB-DIF was assessed by comparing its mean value to that of the control condition under varying levels of contamination. All tests of statistical significance were conducted at a significance level of 0.05.
Results

Table 7.1 shows the performance of MB-DIF for sample sizes of 1000 when the means of the group ability distributions were equal. The first condition reported (having zero contamination) is a control condition in which no contamination was contributed by non-studied items. Consider first the Type I error rates of MB-DIF, presented in Table 7.1 under the condition in which no DIF was introduced into the studied item \( (t_c = 0.0) \). The results indicate that the probability of Type I error is approximately 0.03 over all conditions. This probability appears to be unaffected by increased levels of contamination.

The power of MB-DIF to detect DIF under the condition in which DIF was introduced into the studied item \( (t_c = 0.6) \) was also largely robust to the degree of contamination. Under the control condition the mean value of MB-DIF was 0.103, with an associated power of 0.96 (96% of items containing DIF were flagged as significant). As the contamination increased to the maximal level considered, the power remained at 0.96, and the mean value of MB-DIF decreased on slightly to 0.099.

Table 7.2 displays the results for MB-DIF when groups had 250 members each, and the means of the group ability distributions were equal. Unlike the case of 1000 members per group, the probability of a Type I error inflated to a maximum of 0.11 under the condition of 20% of the non-studied items containing DIF. Under lower degrees of contamination, the probability of a Type I error remained at the nominal level of 0.05. The power of MB-DIF decreased only slightly as contamination increased. The most severe decrease in power was observed under the most extreme contamination conditions, for which power was 0.74 compared with 0.78 in the control condition.
Table 7.1
Mean MB-DIF Values, Type I Error Rates, and Power for Sample Sizes of 1000, and Group Ability Distributions Having Equal Means

<table>
<thead>
<tr>
<th>%</th>
<th>$t$</th>
<th>Mean</th>
<th>Type I</th>
<th>Mean</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>-0.000</td>
<td>0.02</td>
<td>0.103</td>
<td>0.96</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>0.000</td>
<td>0.03</td>
<td>0.105</td>
<td>0.95</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
<td>-0.002</td>
<td>0.03</td>
<td>0.101</td>
<td>0.96</td>
</tr>
<tr>
<td>20</td>
<td>0.3</td>
<td>-0.003</td>
<td>0.03</td>
<td>0.099</td>
<td>0.95</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>-0.004</td>
<td>0.03</td>
<td>0.103</td>
<td>0.97</td>
</tr>
<tr>
<td>10</td>
<td>0.6</td>
<td>-0.001</td>
<td>0.02</td>
<td>0.103</td>
<td>0.96</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>-0.004</td>
<td>0.03</td>
<td>0.099</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Note. % refers to the number of contaminated items, $t$ represents the magnitude of DIF introduced into the contaminated items, and $t_s$ refers to the magnitude of DIF introduced into the studied item.

Table 7.2
Mean MB-DIF Values, Type I Error Rates, and Power for Sample Sizes of 250, and Group Ability Distributions Having Equal Means

<table>
<thead>
<tr>
<th>%</th>
<th>$t$</th>
<th>Mean</th>
<th>Type I</th>
<th>Mean</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>0.001</td>
<td>0.03</td>
<td>0.104</td>
<td>0.78</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>-0.002</td>
<td>0.04</td>
<td>0.101</td>
<td>0.76</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
<td>-0.005</td>
<td>0.06</td>
<td>0.100</td>
<td>0.74</td>
</tr>
<tr>
<td>20</td>
<td>0.3</td>
<td>-0.009</td>
<td>0.11</td>
<td>0.098</td>
<td>0.74</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>-0.004</td>
<td>0.04</td>
<td>0.102</td>
<td>0.75</td>
</tr>
<tr>
<td>10</td>
<td>0.6</td>
<td>-0.005</td>
<td>0.05</td>
<td>0.100</td>
<td>0.75</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>-0.007</td>
<td>0.09</td>
<td>0.097</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Note. See Table 7.1 for a description of the headers.
Tables 7.3 and 7.4 display the performance of MB-DIF when samples contained 1000 and 250 members, and the means of the reference and focal group ability distributions were unequal. In general, the trends of these results parallel those of the conditions in which the means of the group ability distributions were equal. For sample sizes of 1000, Type I error rates remained at the nominal level of 0.05, and power was unaffected by increasing contamination, ranging between 0.82 and 0.84 for all conditions. For sample sizes of 250, Type I error rates increased as contamination increased, reaching a probability of 0.19 when 20% of the test items contained DIF. The related power decreased slightly to 0.57 under high levels of contamination compared with the control value of 0.62.

Table 7.3
Mean MB-DIF Values, Type I Error Rates, and Power for Sample Sizes of 1000, and Group Ability Distributions Having Unequal Means

<table>
<thead>
<tr>
<th>%</th>
<th>t</th>
<th>Mean</th>
<th>Type I</th>
<th>Mean</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>0.001</td>
<td>0.06</td>
<td>0.092</td>
<td>0.82</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>0.001</td>
<td>0.05</td>
<td>0.094</td>
<td>0.84</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
<td>-0.002</td>
<td>0.06</td>
<td>0.094</td>
<td>0.84</td>
</tr>
<tr>
<td>20</td>
<td>0.3</td>
<td>-0.003</td>
<td>0.05</td>
<td>0.091</td>
<td>0.82</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>0.000</td>
<td>0.05</td>
<td>0.096</td>
<td>0.84</td>
</tr>
<tr>
<td>10</td>
<td>0.6</td>
<td>-0.002</td>
<td>0.04</td>
<td>0.092</td>
<td>0.82</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>-0.006</td>
<td>0.05</td>
<td>0.088</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Note. See Table 7.1 for a description of the headers.
Table 7.4
Mean MB-DIF Values, Type I Error Rates, and Power for Sample Sizes of 250, and Group Ability Distributions Having Unequal Means

<table>
<thead>
<tr>
<th>%</th>
<th>t</th>
<th>Mean</th>
<th>Type I</th>
<th>Mean</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>-0.002</td>
<td>0.05</td>
<td>0.096</td>
<td>0.62</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>-0.003</td>
<td>0.06</td>
<td>0.088</td>
<td>0.58</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
<td>0.000</td>
<td>0.10</td>
<td>0.090</td>
<td>0.60</td>
</tr>
<tr>
<td>20</td>
<td>0.3</td>
<td>-0.008</td>
<td>0.19</td>
<td>0.087</td>
<td>0.61</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>-0.001</td>
<td>0.06</td>
<td>0.093</td>
<td>0.60</td>
</tr>
<tr>
<td>10</td>
<td>0.6</td>
<td>-0.005</td>
<td>0.08</td>
<td>0.092</td>
<td>0.62</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>-0.008</td>
<td>0.18</td>
<td>0.084</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Note. See Table 7.1 for a description of the headers.

Discussion

The results suggest that MB-DIF is largely robust to the presence of contamination, being virtually unaffected when sample sizes are large, and mildly affected when sample sizes are small. These findings indicate that MB-DIF provides an effective alternative to other DIF statistics, having the advantage that it controls for the effects of CIM.

Comparing the performance of MB-DIF to that of the MH adjustment (see Chapter Five), we find that MB-DIF was consistently more robust to the effects of CIM than the MH adjustment. When group size equaled 1000, the power of MB-DIF was virtually unaffected by CIM, compared with a decrease in power of up to 0.06 of the adjusted MHχ². When sample sizes were small (N = 250), the decrease in power attributable to CIM for MB-DIF tended to be on the order of half as large as that for the adjusted MHχ². However, MB-DIF displayed Type I error rates that were up to twice as large as those of the adjusted MHχ² when group ability distributions had unequal means and sample sizes were small (N = 250).
Despite the apparent robustness of \textit{MB-DIF} to the effects of CI\textsubscript{M}, there are several limitations of \textit{MB-DIF}. These concern: (1) which items to adjust for in the calculation of the focal group probability of success on the target ability dimension (see Equation 6.2a), (2) the statistical assumptions made in the development of the theory of the parameter estimation (see Chapters Three and Four), and (3) the generalizability of the results to non-simulated data. These limitations were previously discussed in Chapter Five for the MH adjustment, and are thus not discussed further here. It should be noted, however, that the consequences of these limitations are equally applicable to \textit{MB-DIF}.

An additional limitation of \textit{MB-DIF} is the inflated Type I error rate when sample sizes were small \((N = 250)\) and contamination was high. These high error rates can be attributed to the lack of stability of the estimates of manifest bias at each observed score level (see Equation 6.2b) when there are few examinees at each score level. With a sample of only 250 across 41 possible score categories, very few of the score categories will contain more than 20 combined reference and focal group members. This situation not only affects the stability of the difference in observed probability of correct response between the groups, but almost certainly causes the estimation of \(P(\Theta = 1|G = F, m = j)\) to be unstable. Since the procedure used to estimate \(P(\Theta = 1|G = F, m = j)\) constrains it to be higher than the observed focal group item difficulty at observed score category \(j\) (because it is assumed that bias can never act to increase the probability of a focal group correct response), the instability of the estimation when sample size is small can cause the estimate of the focal group probability of success on the target ability dimension at each observed score category to be spuriously high, but not spuriously low. This situation is exacerbated when reference and focal group
members are sampled from ability distributions having different means, in which case there will rarely be large numbers from both groups in the same score category. As a result, if contamination is suspected to be large, MB-DIF should be limited to cases in which sample sizes are greater than 250. Further research is required to establish the Type I error rates when sample sizes are between 250 and 1000.
Chapter Eight

Practical Implications of Findings

Summary of the Major Findings

There are three major findings of this body of work. First, CIM acts to decrease the magnitude of $MHD-DIF$ and the power of the $MH\chi^2$ in detecting DIF. The most severe levels of contamination examined here led to: (1) a mean deviation of the $MHD-DIF$ from control values of up to 0.33 units, (2) a decrease in probability of being flagged as an ETS Category C item of up to 60%, and (3) an increase in the percentage of Type I errors of the $MH\chi^2$ of up to 400%. The second finding is that the effects of contamination on the MH procedure can be controlled for by a generalized binomial adjustment. This adjustment consistently corrects for 50% to 80% of the CIM-induced underestimation in the mean $MHD-DIF$ values, and 70% to 100% of the CIM-induced decrease in the number of items flagged as ETS Category C. The third major finding pertains to the development of an alternative DIF detection statistic ($MB-DIF$). $MB-DIF$ proved to be robust to CIM when sample sizes were large ($N = 1000$), but was mildly affected by CIM when sample sizes were small ($N = 250$). $MB-DIF$ was consistently more robust to CIM than the MH adjustment.

Practical Implications

This research has several implications for practical test development procedures. First, the results show that the effects of CIM must be considered when assessing the magnitude of DIF. The issue of matching criterion contamination is typically addressed in
test development by removing items with high levels of DIF (e.g., Category C items) from the test and rerunning the MH procedure for each of the remaining items. However, when several items contain only moderate levels of DIF, and thus are not removed, substantial matching criterion contamination will still lead to a substantial decrease in the power of the MH procedure.

The effects of CIM on the distribution of $MH\chi^2$ has critical implications concerning simulation research in which more than one item on the simulated test contains bias. In such cases, the $MH\chi^2$ is no longer distributed as a chi-square variable with one degree of freedom. Indeed, this study showed that under severe contamination the Type I error rate of the $MH\chi^2$ increased to over 0.2 when the intended nominal Type I error rate was 0.05. As a result, simulation research making use of the $MH\chi^2$ must consider the effects of contamination on the distribution of the $MH\chi^2$.

The results of Chapters Five and Six indicate that it is possible either to adjust the MH value to counteract the effects of CIM, or use an alternative statistic that is robust to the effects of CIM (e.g., MB-DIF). Using either of these procedures, the ability of test developers to detect biased items will be enhanced, thus improving the validity of test scores.

Limitations to Practical Application

Several limitations of the procedures developed and assessed in this study may affect their practical application to test development. The most imposing of these limitations is the computational complexity of the methods of adjusting for CIM presented in Chapter Four.
This complexity has two primary sources: (1) determining which items to adjust for, and (2) performing the calculations required for the relevant parameter estimations. It is possible that the costs associated with these complications outweigh the benefits achieved with the utilization of such procedures.

There are several possible ways of addressing the above limitation. First, developing software to conduct the computations of the MH adjustment and MB-DIF would solve the problem of the high computational demands. Computer programs for such procedures have been developed in SAS for the purpose of the simulation research presented here, and can be adapted for use in applied test development procedures. A second solution to the limitation of computational complexity is to develop other criteria that are less computationally intensive. For example, DIF analysts may wish to modify their criteria for flagging potentially biased items, making them more liberal as the number of items containing moderate or high levels of DIF increases. While such procedures would not make use of the alternative statistical procedures developed here, they would at least address the problem of a decrease in power of DIF detection as contamination increases.

A second limitation related to that discussed above concerns distinguishing between those items containing DIF and those items containing bias. The presence of DIF does not necessarily imply the existence of bias. As a consequence, it may be a mistake to adjust the MH value for all items containing significant levels of DIF, since items containing significant levels of DIF may not contain bias, and thus may not be causing focal group examinees to be classified systematically below their valid score. Thus, there is an inherent circularity imbedded within the methods proposed for the MH adjustment and MB-DIF; we seek to
adjust for items that contain bias (not DIF), but if we prove that such items do indeed contain bias, then we would be likely to remove such items from the test altogether, in which case there would exist no harmful effects of CIM on the validity of the matching criterion.

This limitation has two components, and I will address each individually. The first component states that if DIF is not necessarily bias, and if we are adjusting for contamination caused by DIF, then we are adjusting for something we do not necessarily want to adjust for. Since the MH adjustment assumes all DIF is bias, it in effect offers an upper bound to the valid ratio of reference and focal group odds of success. Similarly, MB-DIF offers an upper bound to the valid difference between the reference and focal group item-test regressions. Since the adjusted MH value and MB-DIF are estimated upper bounds to the magnitude of DIF, they can be used to alert test developers of how large the estimate of DIF could be if all of the DIF in the studied and non-studied items is attributable to a systematic invalidity in the test.

The second component of the limitation concerns not having the need for an adjustment if we are able to identify all of the biased items on the test. I respond to this by noting that it is often the case that items containing small or moderate levels of bias are retained in the test because their DIF levels are not high enough to warrant their removal without compromising the test's reliability and validity. In this case, the adjustment procedures proposed here become immediately relevant in offering a means to include the questionable items, without having the estimate of DIF in the studied item become systematically negatively biased.

A final limitation, which has been discussed in Chapters Five and Six, is the extent to
which the assumptions made in developing the MH adjustment and MB-DIF are met in real testing data. The effects of violating these assumptions are currently unknown. Future research might address this issue through investigating the performance of these statistics using simulated data in which these assumptions are not met.
References


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


Shealy, R. T., & Stout, W. F. (1993b). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 54*, 159-194.


Appendix A

Simulation Program for Study One

DATA SIM201B;

*ANGLES==================================================================;
*SETTING CONSTANTS:
*ANGLES==================================================================;

*NUMITEM IS THE NUMBER OF ITEMS ON THE TEST;
*THETA IS THE DIFFERENCE BETWEEN THE MEAN OF THE ABILITY DISTRIBUTIONS OF THE
REFERENCE AND FOCAL GROUPS;
*GRP EQUALS THE CONDITION;

NUMITEM = 20;
THETA = 0;
GRP = 0;
COMMA=',';

FILENAME OUT 'C:\RDP\MHVALUES'.LRECL=100;
FILE OUT;

*ANGLES==================================================================;
*HERE ARE THE FACTORS VARIED IN THE STUDY:
*ANGLES==================================================================;

*MAGS = MAGNITUDE OF DIF IN STUDIED ITEM;
*NCI = NUMBER OF CONTAMINATED ITEMS;
*MAGNS = MAGNITUDE OF DIF IN CONTAMINATED ITEMS;

DO MAGS = 1 TO 3;
DO NCI = 1 TO 3;
DO MAGNS = 1 TO 4;

*DETERMINE THE MAGNITUDE OF DIF IN THE STUDIED ITEM (DS) FOR THIS CONDITION;
IF MAGS = 1 THEN DS = 0;
ELSE IF MAGS = 2 THEN DS = .40;
ELSE DS = .60;

*DETERMINE THE NUMBER OF CONTAMINATING ITEMS (NUMCONT) FOR THIS CONDITION;
IF NCI = 1 THEN NUMCONT = 1;
ELSE IF NCI = 2 THEN NUMCONT = 2;
ELSE NUMCONT = 4;

*DETERMINE THE MAGNITUDE OF DIF IN THE NON-STUDIED ITEMS (DNS) FOR THIS
CONDITION;
IF MAGNS = 1 THEN DNS = 0;
ELSE IF MAGNS = 2 THEN DNS = .20;
ELSE IF MAGNS = 3 THEN DNS = .40;
ELSE DNS = .60;

*////////////////////////////////////////////////////////////////////////////////////////////////////////;
*SETTING THE ETS CATEGORY COUNTERS TO ZERO;
*////////////////////////////////////////////////////////////////////////////////////////////////////////;

ETSOBA = 0;
ETSOBB = 0;
ETSOBC = 0;

*DESIGNATE THE CONDITION BY 'GRP':
GRP = GRP + 1;

*////////////////////////////////////////////////////////////////////////////////////////////////////////;
*START OF THE 1000 TRIALS TO BE RUN;
*////////////////////////////////////////////////////////////////////////////////////////////////////////;

DO G = 1 TO 1000;

*////////////////////////////////////////////////////////////////////////////////////////////////////////;
*LIST ALL ARRAYS;
*////////////////////////////////////////////////////////////////////////////////////////////////////////;

ARRAY TOTOBS{21} TOTOBS0-TOTOBS20;
ARRAY AOBS{21} AOBS0-AOBS20;
ARRAY BOBS{21} BOBS0-BOBS20;
ARRAY COBS{21} COBS0-COBS20;
ARRAY DOBS{21} DOBS0-DOBS20;

ARRAY NUMOBS{21} NUMOBS0-NUMOBS20;
ARRAY DENOBS{21} DENOBS0-DENOBS20;

ARRAY K1OBS{21} K1OBS0-K1OBS20;
ARRAY K2OBS{21} K2OBS0-K2OBS20;
ARRAY K3OBS{21} K3OBS0-K3OBS20;

ARRAY PROB{20} PROB1-PROB20;
ARRAY OBS{20} OBS1-OBS20;

ARRAY Y{21} Y1-Y21;
ARRAY A{21} A1-A21;
ARRAY APRE{21} APRE1-APRE21;
ARRAY B{21} B1-B21;

ARRAY PRATM{21} PRATM0-PRATM20;
ARRAY NRATM{21} NRATM0-NRATM20;
ARRAY NFATM{21} NFATM0-NFATM20;
ARRAY D1A{21} D1A0-D1A20;
ARRAY D2A{21} D2A0-D2A20;

ARRAY ABOBS{21} ABOBS0-ABOBS20;
ARRAY ACOBS{21} ACOBS0-ACOBS20;
ARRAY CDOBS[21] CDOBS0-CDOBS20;
ARRAY BDOBS[21] BDOBS0-BDOBS20;
ARRAY ALLOBS[21] ALLOBS0-ALLOBS20;
ARRAY EXPA[21] EXPA0-EXPA20;
ARRAY VARA[21] VARA0-VARA20;

*SET ALL COUNTERS WITHIN A CONDITION TO ZERO;

DO J = 1 TO 21:
  TOTOBS(J) = 0;
  AOBS(J) = 0;
  BOBS(J) = 0;
  COBS(J) = 0;
  DOBS(J) = 0;
  NUMOBS(J) = 0;
  DENOBS(J) = 0;
END;

*DETERMINE THE PROBABILITY OF SUCCESS FOR EACH PERSON;

*HERE WE ARE GOING TO SAMPLE TWO IRT PARAMETERS A AND B;
*A IS THE DISCRIMINATION PARAMETER OF THE IRT MODEL;
*B IS THE DIFFICULTY PARAMETER OF THE IRT MODEL;
*A IS SAMPLED FROM A LOG-NORMAL DISTRIBUTION WHERE A = EXP(Z), AND Z~N(0,.1225);
*B IS SAMPLED FROM N(0,1);

DO L = 1 TO NUMITEM;

*DRAW TWO N(0,1) VARIATES;
B{L}=RANNOR(0);
APRE{L}=RANNOR(0);

*SET ONE N(0,1) VARIATE TO N(0,.1225) AND THEN TRANSFORM IT TO EXP[N(1,.1225)];
APRE{L} = APRE{L}*(.1225);
A{L} = EXP(APRE{L});
END;

*GENERATE 2000 TEST RESPONSES, 1000 FOR EACH GROUP;
DO N=1 TO 2000;

*SAMPLE THETA VALUES FROM N(0, 1);
X=RANNOR(0);

*FOR EACH ITEM, OBTAIN A UNIFORM VARIATE;
DO L = 1 TO 20;
Y{L}=RANUNI(0);
END;
*DETERMINE REFERENCE (2) AND FOCAL (1) GROUP MEMBERSHIP GROUP MEMBERSHIP:*
IF N < 1001 THEN GROUP = 1;
ELSE GROUP = 2;

*DETERMINE THETA VALUE:*
IF N < 1001 THEN X = X - THETA;
ELSE X = X;

*DETERMINE PROBABILITY OF CORRECT RESPONSE FOR CONTAMINATING ITEMS:*
DO I = 1 TO NUMCONT;
IF GROUP = 1 THEN PROB(I) = .2 + (.8)/(1 + (EXP((-1.7)*(A(I))*(X - B(I)) - DNS))));
ELSE PROB(I) = .2 + (.8)/(1 + (EXP((-1.7)*(A(I))*(X - B(I)))))
END;

*DETERMINE PROBABILITY OF CORRECT RESPONSE FOR NON-CONTAMINATING ITEMS:*
DO I = (NUMCONT + 1) TO (NUMITEM - 1):
PROB(I) = .2 + (.8)/(1 + (EXP((-1.7)*(A(I))*(X - B(I)))))
END;

*DETERMINE PROBABILITY OF CORRECT RESPONSE FOR STUDIED ITEM:*
IF GROUP = 1 THEN PROB(20) = .2 + (.8)/(1 + (EXP((-1.7)*(A(20))*(X - B(20)) - DS))));
ELSE PROB(20) = .2 + (.8)/(1 + (EXP((-1.7)*(A(20))*(X - B(20)))))
END;

*DETERMINE OBSERVED SCORE FOR EACH ITEM:*
DO I = 1 TO NUMITEM:
IF Y(I) <= PROB(I) THEN OBS(I) = 1;
ELSE OBS(I) = 0;
END;

*DETERMINE THE TOTAL TEST SCORE:*
OBSCORE = SUM (OF OBS1- OBS20);

*HERE WE WANT TO DETERMINE THE NUMBER OF OBSERVATIONS IN EACH CELL OF A TWO-BY-TWO TABLE AT EACH SCORE CATEGORY, WHERE THE VARIABLES Crossed ARE GROUP MEMBERSHIP (FOCAL VS. REFERENCE) AND ITEM PERFORMANCE (CORRECT VS. INCORRECT). THESE CELLS ARE CODES AS 1, 2, 3, AND 4 FOR REFCOR, REFINCOR, FOCCOR, FOINCOR:*
IF GROUP = 2 AND OBS20 = 1 THEN OBSCAT = 1;
ELSE IF GROUP = 2 AND OBS20 = 0 THEN OBSCAT = 2;
ELSE IF GROUP = 1 AND OBS20 = 1 THEN OBSCAT = 3;
ELSE OBSCAT = 4;

*A ROLLING TALLY OF THE NUMBER OF OBSERVATIONS IN EACH OF THE 4 CELLS DESCRIBED ABOVE IN KEPT. HERE, THE CELLS CODED 1, 2, 3, AND 4 ARE DENOTED BY 'AOBS', 'BOBS', 'COBS', 'DOBS'. IN ADDITION, THE TOTAL NUMBER ACROSS ALL CELLS IS DENOTED BY 'TOTOLBS'. THESE COUNTERS ARE CONDUCTED FOR EACH SCORE CATEGORY 'J':*
J = OBSCORE + 1;
TOTOLS(J) = TOTOLS(J) + 1;
IF OBSCAT = 1 THEN AOBS(J) = AOBS(J) + 1;
ELSE IF OBSCAT = 2 THEN BOBS(J) = BOBS(J) + 1;
ELSE IF OBSCAT = 3 THEN COBS(J) = COBS(J) + 1;
ELSE DOBS{J} = DOBS{J} + 1;

END;

*-------------------------------------------------------------------------------------------------------;
*CALCULATION OF THE OBSERVED MANTEL-HAENSZEL;
*-------------------------------------------------------------------------------------------------------;

*COMPUTE THE NUMERATOR AND DENOMINATOR OF THE MH COMMON ODDS RATIO:
DO J = 1 TO 21:
NUMOBS {J} = (AOBS {J} • DOBS {J})/ TOTOB {J};
DENOBS {J} = (BOBS {J} • COBS {J})/ TOTOB {J};
END;

*COMPUTE THE MH COMMON ODDS RATIO (MHOBS):
NUMERO = SUM ((NUMOBS0-NUMOBS20);
DENOMO = SUM ((DENOBS0-DENOBS20);
MHOBS = NUMERO / DENOMO;

*COMPUTE THE MH LOG ODDS RATIO (LOROBS) AND MHD-DIF INDEX (DDIFOBS):
LOROBS = LOG(MHOBS);
DDIFOBS = -2.35 • LOROBS;

*CALCULATION OF THE STANDARD ERROR FOR THE OBSERVED MANTEL-HAENSZEL LOR AND MHD-DIF INDEX;
DO J = 1 TO 21:
K1OBS {J} = ((AOBS {J} • DOBS {J}) +(MHOBS • BOBS {J} • COBS {J} )/(TOTOB {J} **2);
K2OBS {J} = (AOBS {J} + DOBS {J} + (MHOBS • (BOBS {J} + COBS {J} ) );
K3OBS {J} = K1OBS {J} * K2OBS {J};
END;
K4OBS = SUM (OF K3OBS1-K3OBS21);
SDLOROBS = SQRT ((1/(2*NUMERO**2))*K4OBS);
SDDDIFOX = 2.35*SDLOROBS;

*-------------------------------------------------------------------------------------------------------;
*DETERMINING THE ETS CLASSIFICATION;
*-------------------------------------------------------------------------------------------------------;

*COMPUTE THE ABSOLUTE VALUE OF MHD-DIF;
OABSDIFF=ABS(DDIFOBS);

*COMPUTE THE Z-SCORE OF THE ABSOLUTE VALUE OF MHD-DIF;
OZSCORE = ABS(DDIFOBS/SDDDIFOX);

*TEST SIGNIFICANCE FROM 1 OF MHD-DIF ABSOLUTE VALUE;
OBSDIF1 = ABS(OABSDIFF - 1)/SDDDIFOX;

*DETERMINE ETS CATEGORY (ETSCATOB);
IF OABSDIFF < 1 OR OZSCORE <= 2 THEN
ETSCATOB = 1;
ELSE IF OBSDIFF >= 1.5 AND OBSDIFF1 > 2 THEN
ETSCATOB = 3;
ELSE ETSCATOB = 2;

*KEEPING COUNT OF THE NUMBER OF ITEMS IN EACH ETS CATEGORY. CATEGORIES A, B, AND C ARE DENOTED BY 'ETSOBA', 'ETSOBB' AND 'ETSOBC':

IF ETSCATOB = 1 THEN ETSOBA = 1;
ELSE ETSOBA = 0;

IF ETSCATOB = 2 THEN ETSOBB = 1;
ELSE ETSOBB = 0;

IF ETSCATOB = 3 THEN ETSOBC = 1;
ELSE ETSOBC = 0;

*CALCULATION OF CHI-SQUARE FOR MH;

DO J = 1 TO 21;
ABOBS{J} = AOBS{J} + BOBS{J};
ACOBS{J} = AOBS{J} + COBS{J};
CDOBS{J} = COBS{J} + DOBS{J};
BDABS{J} = BOBS{J} + DOBS{J};
ALLOBS{J} = AOBS{J} + BOBS{J} + COBS{J} + DOBS{J};

*COMPUTE THE EXPECTED VALUE OF THE NUMBER IN CELL A;
EXPA{J} = ((ACOBS{J}) * (ABOBS{J})) / ALLOBS{J};

*COMPUTE THE VARIANCE OF THE NUMBER IN CELL A;
VARA{J} = ((ABOBS{J}) * (ACOBS{J}) * (CDOBS{J}) * (BDABS{J})) / ((ALLOBS{J} ** 2) * (ALLOBS{J} - 1));
END:

*SUM THE NUMBER OF OBSERVATIONS IN CELL A (TOTA), THE VARIANCE OF THE NUMBER IN A (TVARA), AND THE EXPECTED VALUE OF THE NUMBER IS CELL A (TEXPA) ACROSS ALL SCORE CATEGORIES:
TOTA = SUM (OF AOBS0-AOBS20);
TVARA = SUM (OF VARA0-VARA20);
TEXPA = SUM (OF EXPA0-EXPA20);

*COMPUTING THE MH CHI-SQUARE VALUE (MHCHI);
MHCHI = ((ABS(TOTA - TEXPA) - .5)**2) / TVARA;

*DETERMINING THE SIGNIFICANCE OF THE MH CHI-SQUARE VALUE (SIG);
IF MHCHI > 3.84 THEN SIG = 1;
ELSE SIG = 0;

OUTPUT;

*WRITING THE RESULTS TO A COMMA-DELIMITED FILE;

103
PUT @1 MHOBS 7.5
@11 COMMA S1.
@13 GRP 2:;

END;
END;
END;
END;

*COMPUTE MEAN OF MHD-DIF, POWER OF MH CHI-SQUARE
AND PROPORTION IN EACH ETS CATEGORY;
*

PROC MEANS MEAN STD SKEWNESS KURTOSIS;
BY GRP;
VAR GRP DS DNS NUMCONT MHOBS DDIFOBS ETSOBA ETSOBB ETSOBC SIG;

RUN:

104
Appendix B

Simulation Program for Study Two

DATA SIM201B:

*=======================================================================;
*CONSTANTS;
*=======================================================================;

NUMITEM = 40;
NUMEX = 2000;
NUMGRP = NUMEX/2;
THETA = 0;
GRP = 0;
DS = .6;
NUMCONT = 8;
DNS = .6:

*NUMITEM IS THE NUMBER OF ITEMS ON THE TEST;
*NUMEX IS THE TOTAL NUMBER OF EXAMINEES;
*NUMGRP IS THE NUMBER OF EXAMINEES IN EACH GROUP;
*THETA IS THE DIFFERENCE IN MEAN OF THE REFERENCE AND FOCAL ABILITY
DISTRIBUTIONS;
*GRP IS THE CONDITION;
*DS IS THE DIFF IN THE STUDIED ITEM;
*NUMCONT IS THE NUMBER OF CONTAMINATING ITEMS;
*DMS IS THE DIFF IN THE NON-STUDIED ITEMS;

*=======================================================================;
*SETTING THE ETS CATEGORY COUNTERS TO ZERO;
*=======================================================================;

ETSOBA = 0;
ETSOBB = 0;
ETSOBC = 0;

*STARTING THE LOOP FOR EACH OF THE TRIALS;
DO G = 1 TO 1000;

*=======================================================================;
*LIST ALL ARRAYS;
*=======================================================================;

ARRAY TOTOB{41} TOTOBS1-TOTOBS41;
ARRAY AOB{41} AOB1-AOB41;
ARRAY BOBS{41} BOBS1-BOBS41;
ARRAY COBS{41} COBS1-COBS41;
ARRAY DOBS[41] DOBS1-DOBS41;

ARRAY NUMOBS[41] NUMOBS1-NUMOBS41;
ARRAY DENOBS[41] DENOBS1-DENOBS41;

ARRAY K1OBS[41] K1OBS1-K1OBS41;
ARRAY K2OBS[41] K2OBS1-K2OBS41;
ARRAY K3OBS[41] K3OBS1-K3OBS41;
ARRAY PROBB[41] PROB1-PROB41;
ARRAY OBS[41] OBS1-OBS41;
ARRAY Y[41] Y1-Y41;
ARRAY AA[41] AA1-AA41;
ARRAY APRE[41] APRE1-APRE41;
ARRAY BB[41] BB1-BB41;
ARRAY PRATM[41] PRATM1-PRATM41;
ARRAY NRATM[51] NRATM1-NRATM51;
ARRAY NFATM[51] NFATM1-NFATM51;
ARRAY D1A[41] D1A1-D1A41;
ARRAY D2A[41] D2A1-D2A41;

ARRAY ABOBS[41] ABOBS1-ABOBS41;
ARRAY ACOBS[41] ACOBS1-ACOBS41;
ARRAY CDOBS[41] CDOBS1-CDOBS41;
ARRAY BDOBS[41] BDOBS1-BDOBS41;
ARRAY ALLOBS[41] ALLOBS1-ALLOBS41;
ARRAY EXPAC[41] EXPAC1-EXPAC41;
ARRAY VARAC[41] VARAC1-VARAC41;

ARRAY ABCOR[41] ABCOR1-ABCOR41;
ARRAY ACCOR[41] ACCOR1-ACCOR41;
ARRAY CDCOR[41] CDCOR1-CDCOR41;
ARRAY BDCOR[41] BDCOR1-BDCOR41;
ARRAY ALLCOR[41] ALLCOR1-ALLCOR41;
ARRAY EXPAC[41] EXPAC1-EXPAC41;
ARRAY VARAC[41] VARAC1-VARAC41;

ARRAY TTPROB[9] TTPROB1-TTPROB9;
ARRAY DELTA[9] DELTA1-DELTA9;
ARRAY NUMTR[9] NUMTR1-NUMTR9;
ARRAY WACK[9] WACK1-WACK9;

ARRAY PR[9, 51] PR1-PR459;
ARRAY B[9, 41] B1-B369;
ARRAY C[9, 41] C1-C369;
ARRAY D[9, 41] D1-D369;
ARRAY D1NS{9,41} D1NS1-D1NS369;
ARRAY D2NS{9,41} D2NS1-D2NS369;
ARRAY TTN{9,41} TTN1-TTN369;
ARRAY TOT{9,41} TOT1-TOT369;
ARRAY PRB{9,41} PRB1-PRB369;
ARRAY D3NS{13} D3NS1-D3NS13;
ARRAY D4NS{13} D4NS1-D4NS13;
ARRAY DELTNS{13} DELTNS1-DELTNS13;
ARRAY IT{13} IT1-IT13;
ARRAY CAT{13} CAT1-CAT13;
ARRAY PREB{41} PREB1-PREB41;
ARRAY TTCOR{41} TTCOR1-TTCOR41;
ARRAY TTPR{41} TTPR1-TTPR41;
ARRAY OKB{41} OKB1-OKB41;
ARRAY OKC{41} OKC1-OKC41;
ARRAY CORPR{41} CORPR1-CORPR41;
ARRAY ACOR{41} ACOR1-ACOR41;
ARRAY BCOR{41} BCOR1-BCOR41;
ARRAY NUMCOR{41} NUMCOR1-NUMCOR41;
ARRAY DENCOR{41} DENCOR1-DENCOR41;
ARRAY NFMU{81} NFMU1-NFMU81;
ARRAY K1COR{41} K1COR1-K1COR41;
ARRAY K2COR{41} K2COR1-K2COR41;
ARRAY K3COR{41} K3COR1-K3COR41;
ARRAY ORIGN{41} ORIGN1-ORIGN41;
ARRAY PRM{9,51} PRM1-PRM459;
ARRAY PFM{9,51} PFM1-PFM459;
ARRAY DIFM{9,41} DIFM1-DIFM369;
ARRAY DELA{9,41} DELA1-DELA369;
ARRAY DELB{9,41} DELB1-DELB369;
ARRAY DELC{9,41} DELC1-DELC369;
ARRAY SUMA{9} SUMA1-SUMA9;
ARRAY DELCHA{9} DELCHA1-DELCHA9;
ARRAY SUMB{9} SUMB1-SUMB9;

*////////////////////////////////////////////////////////////////////////////////////////;
*SET ALL COUNTERS TO ZERO;
*////////////////////////////////////////////////////////////////////////////////////////;

DO J = 1 TO 41;
TOTOBS{J} = 0;
AOBS{J} = 0;
BOBS{J} = 0;
COBS{J} = 0;
DOBS{J} = 0;
NUMOBS{J} = 0;
DENOBS{J} = 0;
DO J = 1 TO 9;
TTPROB(J) = 0;
D3NS(J) = 0;
D4NS(J) = 0;
END;

DO J = 1 TO 9;
DO K = 1 TO 41;
PRB(J, K) = 0;
TOT(J, K) = 0;
A(J, K) = 0;
B(J, K) = 0;
C(J, K) = 0;
D(J, K) = 0;
END;
END:

*SIMULATING ITEM RESPONSES FOR EACH PERSON;
*SAMPLE DISCRIMINATION (A) AND DIFFICULTY (BB) PARAMETERS;
DO L = 1 TO NUMITEM:

*BB IS SAMPLED FROM N(0,1);
BB(L) = RANNOR(0);

*APRE IS SAMPLED FROM A N(0, 1) DISTRIBUTION;
APRE(L) = RANNOR(0);

*AA IS SAMPLED FROM A LOG-NORMAL DISTRIBUTION WHERE AA = EXP(Z), AND Z ~ N(0, 1.225):
APRE(L) = APRE(L) *.1225;
AA(L) = EXP(APRE(L));
END;

*REPPLICATE FOR ‘NUMEX’ EXAMINEES;
DO N=1 TO NUMEX;

*SAMPLE A THETA VALUE ‘X’ FROM N(0, 1);
X = RANNO

*SAMPLE A UNIFORM VARIATE ‘Y’ FOR EACH ITEM;
DO L = 1 TO (NUMITEM + 1);
Y(L) = RANUNI(0);
END;

*ASSIGN EXAMINEES TO FOCAL (1) AND REFERENCE (2) GROUPS;
IF N < (NUMGRP+1) THEN GROUP = 1;
ELSE GROUP = 2;

108
*ADJUST THE FOCAL GROUP THETA VALUES (X) ACCORDING TO ORDERING OF GROUP MEANS;
IF N < (NUMGRP+1) THEN X = X - THETA;
ELSE X = X;

*DETERMINE THE PROBABILITY OF CORRECT RESPONSE ON THE CONTAMINATED ITEMS;
DO I = 2 TO (NUMCONT+1);
IF GROUP = 1 THEN PROBB[I] = .2 + (.8)*(1/(1 + (EXP((-1.7)*(AA[I])*X - BB[I] - DNS))));
ELSE PROBB[I] = .2 + (.8)*(1/(1 + (EXP((-1.7)*(AA[I])*X - BB[I])))�;
END;

*DETERMINE PROBABILITY OF CORRECT RESPONSE FOR UNCONTAMINATED ITEMS;
DO I = (NUMCONT+2) TO (NUMITEM);
PROBB[I] = .2 + (.8)*(1/(1 + (EXP((-1.7)*(AA[I])*X - BB[I])))�;
END;

*DETERMINE THE PROBABILITY OF CORRECT RESPONSE FOR THE STUDIED ITEM;
IF GROUP = 1 THEN PROBB[I] = .2 + (.8)*(1/(1 + (EXP((-1.7)*(AA[I])*X - BB[I] - DNS))));
ELSE PROBB[I] = .2 + (.8)*(1/(1 + (EXP((-1.7)*(AA[I])*X - BB[I])))�;

*DETERMINE THE MANIFEST SCORE ON EACH ITEM;
DO I = 1 TO NUMITEM;
IF Y[I] <= PROBB[I] THEN OBS[I]=1;
ELSE OBS[I]=0;
END;

*CALCULATE THE TEST SCORE;
DO I = 1 TO NUMITEM;
IF OBS[I] = 1 THEN OBS[I] = 0;
ELSE OBS[I] = OBS[I];
END;

OBSCORE = SUM (OF OBS1-OBS60);

*SUMMING THE NUMBER OF OBSERVATIONS IN EACH CATEGORY;
*STUDIED ITEM ONLY;
*HERE WE WANT TO DETERMINE THE NUMBER OF OBSERVATIONS IN EACH CELL OF A TWO-
BY-TWO TABLE AT EACH SCORE CATEGORY, WHERE THE VARIABLES CROSSED ARE GROUP
MEMBERSHIP (FOCAL VS. REFERENCE) AND ITEM PERFORMANCE (CORRECT VS.
INCORRECT). THESE CELLS ARE CODED AS 1, 2, 3, AND 4 FOR REF/COR, REF/INCOR. FOC/COR,
FOC/INCOR;
IF GROUP=2 AND OBS1=1 THEN OBSCAT=1;
ELSE IF GROUP=2 AND OBS1=0 THEN OBSCAT=2;
ELSE IF GROUP=1 AND OBS1=1 THEN OBSCAT=3;
ELSE OBSCAT=4;

*A ROLLING TALLY OF THE NUMBER OF OBSERVATIONS IN EACH OF THE 4 CELLS
DESCRIBED ABOVE IN KEPT. HERE, THE CELLS CODED 1, 2, 3, AND 4 ARE DENOTED BY
'AOBS', 'BOBS', 'COBS', 'DOBS'. IN ADDITION, THE TOTAL NUMBER ACROSS ALL CELLS IS
DENOTED BY 'TOTOBS'. THESE COUNTERS ARE CONDUCTED FOR EACH SCORE CATEGORY 'J':

\[ J = \text{OBSCORE} + 1; \]
\[ \text{TOTOBS}(J) = \text{TOTOBS}(J) + 1; \]
\[ \text{IF OBSCAT} = 1 \text{ THEN AOBS}(J) = \text{AOBS}(J) + 1; \]
\[ \text{ELSE IF OBSCAT} = 2 \text{ THEN BOBS}(J) = \text{BOBS}(J) + 1; \]
\[ \text{ELSE IF OBSCAT} = 3 \text{ THEN COBS}(J) = \text{COBS}(J) + 1; \]
\[ \text{ELSE DOBS}(J) = \text{DOBS}(J) + 1; \]

*///~////////////////////////////////////////////////////////~/////////////////////////////////////////!////////////;

CATEGORIZING SCORES ACCORDING TO GROUP AND ITEM RESPONSE;
*SUMMING THE NUMBER OF OBSERVATIONS IN EACH CATEGORY;
*NON-STUDIED ITEMS ONLY;
*///!l///////llll/////////////////////////////////////////////////////////////////////////////////!/////////////!//~

CATEGORIZING (CAT) RESPONSES ACCORDING TO GROUP AND RESPONSE. THUS, FOR EACH PERSON, THEY ARE CATEGORIZED FOR EACH OF THE K BIASED ITEMS:

\[ \text{IF GROUP}=2 \text{ AND OBS}(K)=1 \text{ THEN CAT}(K)=1; \]
\[ \text{ELSE IF GROUP}=2 \text{ AND OBS}(K)=0 \text{ THEN CAT}(K)=2; \]
\[ \text{ELSE IF GROUP}=1 \text{ AND OBS}(K)=1 \text{ THEN CAT}(K)=3; \]
\[ \text{ELSE CAT}(K)=4; \]

ROLLING COUNT OF THE NUMBER IN EACH CATEGORY FOR EACH BIASED ITEM (K) AT EACH SCORE CATEGORY (J). HERE REF/COR, REF/INCOR, FOC/COR, FOC/INCOR ARE DENOTED BY 'A', 'B', 'C', 'D'. IN ADDITION, THE TOTAL NUMBER OF OBSERVATIONS AT EACH SCORE CATEGORY IS DENOTED BY 'TOT':

\[ J = \text{OBSCORE} + 1; \]
\[ \text{TOT}(K, J) = \text{TOT}(K, J) + 1; \]
\[ \text{IF CAT}(K) = 1 \text{ THEN A}(K, J) = A(K, J) + 1; \]
\[ \text{ELSE IF CAT}(K) = 2 \text{ THEN B}(K, J) = B(K, J) + 1; \]
\[ \text{ELSE IF CAT}(K) = 3 \text{ THEN C}(K, J) = C(K, J) + 1; \]
\[ \text{ELSE D}(K, J) = D(K, J) + 1; \]

END;

END; *THAT ENDS THE LOOP FOR SUBJECTS;

*///~/////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////;

CALCULATING THE MH FOR THE STUDIED ITEM;
*///~/////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////;

\[ \text{NUMERO} = 0; \]
\[ \text{DENOMO} = 0; \]

COMPUTING THE NUMERATOR AND DENOMINATOR OF THE MH COMMON ODDS RATIO;

\[ \text{DO J = 1 TO (NUMITEM + 1);} \]
\[ \text{NUMOBS}(J) = (\text{AOBS}(J) \times \text{DOBS}(J)) / \text{TOTOBS}(J); \]
\[ \text{IF NUMOBS}(J) = . \text{ THEN NUMOBS}(J) = 0; \]
DENOBS {J} = (BOBS {J} * COBS {J}) / TOTOB {J};
IF DENOBS {J} = . THEN DENOBS {J} = 0;

NUMERO = NUMERO + NUMOBS {J};
DENOMO = DENOMO + DENOBS {J};
END:

*COMPUTING THE MH COMMON ODDS RATIO (MHOBS):
MHOBS = NUMERO / DENOMO;

*COMPUTE THE MHD-DIF (DDIFOBS):
LOROBS = LOG(MHOBS);
DDIFOBS = -2.35 * LOROBS;

*////////////////////////////////////////////////////////////////////////////////////////////////////;
*CALCULATION OF THE STANDARD ERROR FOR THE MHD-DIF;
*STUDIED ITEM:
*////////////////////////////////////////////////////////////////////////////////////////////////////;

K4OBS = 0;

DO J = 1 TO (NUMITEM + 1);
K1OBS {J} = ((AOBS {J} * DOBS {J}) + (MHOBS * BOBS {J} * COBS {J}))/((TOOBS {J} ** 2);
K2OBS {J} = (AOBS {J} + DOBS {J}) + (MHOBS * (BOBS {J} + COBS {J})));
K3OBS {J} = K1OBS {J} * K2OBS {J};

IF K3OBS {J} = . THEN K3OBS {J} = 0;
ELSE K3OBS {J} = K3OBS {J};

K4OBS = K4OBS + K3OBS {J};
END;

*STANDARD ERROR FOR MH LOG ODDS RATIO (SDLOROBS);
SDLOROBS = SQRT((1/(2*NUMERO**2)) * K4OBS);

*STANDARD ERROR FOR MHD-DIF (SDDDIFOB);
SDDDIFOB = 2.35 * SDLOROBS;

*////////////////////////////////////////////////////////////////////////////////////////////////////;
*ETS CLASSIFICATION;
*STUDIED ITEM:
*////////////////////////////////////////////////////////////////////////////////////////////////////;

*COMPUTE THE ABSOLUTE VALUE OF MHD-DIF;
OABSDDIF = ABS(DDIFOBS);

*COMPUTE THE ABSOLUTE VALUE OF THE Z-SCORE;
OZSCORE = ABS(DDIFOBS/SDDDIFOB);

*NUMBER OF STANDARD ERRORS FROM ONE;
OBSDIF = ABS(OABSDIF - 1)/SDDDIFOB;
*DETERMINING THE ETS CATEGORY (ETSCATOB);
IF OABSDDIF < 1 OR OZSCORE <=2 THEN
ETSCATOB = 1;
ELSE IF OABSDDIF >= 1.5 AND OBSDIF1 > 2 THEN
ETSCATOB = 3;
ELSE ETSCATOB = 2;

KEEPING A COUNT OF EACH ETS CATEGORY. CATEGORIES A, B, C ARE DENOTED BY
'ETSOBA', ETSOBB', ETSOBC';
IF ETSCATOB = 1 THEN ETSOBA = 1;
ELSE ETSOBA = 0;

IF ETSCATOB = 2 THEN ETSOBB = 1;
ELSE ETSOBB = 0;

IF ETSCATOB = 3 THEN ETSOBC = 1;
ELSE ETSOBC = 0;

*CALCULATION OF MH CHI-SQUARE;
*STUDIED ITEM;
*////////////////////////////////////////////////////////////////////////////////////////////////////////////////;

DO J = 1 TO (NUMITEM + 1);
ABOBS(J) = AOBS(J) + BOBS(J);
ACOBS(J) = AOBS(J) + COBS(J);
CDOBS(J) = COBS(J) + DOBS(J);
BDOBS(J) = BOBS(J) + DOBS(J);
ALLOBS(J) = AOBS(J) + BOBS(J) + COBS(J) + DOBS(J);

*COMPUTE THE EXPECTED VALUE OF THE NUMBER IN CELL A (EXPA);
EXPA(J) = ((ACOBS(J) * (ABOBS(J)))/ALLOBS(J));

*COMPUTE THE VARIANCE OF THE NUMBER IN CELL A (VARA);
VARA(J) = ((ABOBS(J) * (ACOBS(J)) * (CDOBS(J)) * (BDOBS(J))) / ((ALLOBS(J) ** 2) * (ALLOBS(J) - 1)));
END;

*SUM THE TOTAL NUMBER IN CELL A (TOTA), THE EXPECTED NUMBER IN CELL A (TEXPA),
AND THE VARIANCE OF THE NUMBER IN CELL A (TVARA) ACROSS ALL SCORE CATEGORIES:
TOTA = SUM (OF AOBS1-AOBS41);
TVARA = SUM (OF VARA1-VARA41);
TEXPA = SUM (OF EXPAl-EXPA41);

*COMPUTE THE MH CHI-SQUARE FOR THE STUDIED ITEM (MHCHI);
MHCHI = ((ABS(TOTA - TEXPA) - .5) ** 2) / TVARA;

*DETERMINING THE SIGNIFICANCE OF MH CHI-SQUARE (SIG);
IF MHCHI > 3.84 THEN SIG = 1;
ELSE SIG = 0;

*%%% Versions as above stack!***
**CALCULATION OF THE CORRECTED MH VALUES;**
**PROBABILITIES OF SUCCESS FOR NON-STUDIED ITEMS;**

**NUMBERS IN FOCAL (NRATM) AND REFERENCE GROUPS (NRATM):**

DO I = 1 TO (NUMITEM + 1);
NRATM[I] = AOBS[I] + BOBS[I];
NFATM[I] = COBS[I] + DOBS[I];
END;

**PROBABILITY OF SUCCESS FOR THE REFERENCE GROUP (PR) FOR NON-STUDIED ITEMS FOR EACH ITEM (K) AND OBSERVED SCORE CATEGORY (I):**

DO K = 1 TO (NUMCONT + 1);
DO I = 1 TO (NUMITEM + 1);
END;
END;

**SETTING UNDEFINED REFERENCE PROBABILITIES OF SUCCESS TO ZERO FOR EACH STUDIED ITEM (K) AT EACH SCORE CATEGORY (I):**

DO K = 1 TO (NUMCONT+1);
DO I = 1 TO (NUMITEM + 1);
IF NRATM[I] = 0 THEN PR[K, I] = 0;
ELSE PR[K, I] = PR[K, I];
END;
END;

**COMPUTE THE PROBABILITY OF CORRECT RESPONSE FOR THE REFERENCE (PRM) AND FOCAL (PFM) GROUPS FOR EACH BIASED ITEM (K) AT EACH SCORE CATEGORY (I):**

DO K = 1 TO (NUMCONT + 1);
DO I = 1 TO (NUMITEM + 1);
PFM[K, I] = C[K, I]/NFATM[I];

**SETTING UNDEFINED PROBABILITIES OF SUCCESS TO ZERO FOR EACH STUDIED ITEM (K) AT EACH SCORE CATEGORY (I):**

IF PRM[K, I] = . THEN PRM[K, I] = 0;
IF PFM[K, I] = . THEN PFM[K, I] = 0;
END;
END;

**COMPUTING THE PROBABILITY OF SUCCESS ON THE TARGET ABILITY DIMENSION FOR THE FOCAL GROUP:**
*START A LOOP FOR EACH SCORE CATEGORY;
DO I = 1 TO (NLTMITEM + 1);

PREB[I] = 0;
NHIGH[I] = 0;

*START A LOOP FOR EACH NON-STUDIED CONTAMINATING ITEM;
DO BI = 1 TO NUMCONT;

WICKA = 1;
WICKB = 1;
WICKC = 1;

*FIRST, WE DETERMINE THE NUMBER OF COMBINATIONS OF MISSING B ITEMS
OF "NUMCONT" POSSIBLE ITEMS;

* NOTE THAT ‘BI’ = NUMBER OF ITEMS BIASED AGAINST,
‘NUMCOMB’ = NUMBER OF POSSIBLE COMBINATIONS
OF BEING BIASED AGAINST B ITEMS, ‘NUMBIAS’ = NUMBER
OF BIASED NON-STUDIED ITEMS ON THE TEST;

*COMPUTE THE FACTORIAL (NUMBIAS!). HERE, ‘WICKA’ EQUALS THE ‘NUMBIAS’
FACTORIAL;
DO W = 1 TO NUMCONT;
WICKA = WICKA * (NUMCONT - W + 1);
END;

*WICKB EQUALS THE FACTORIAL OF THE NUMBER BIASED AGAINST (BI);
IF BI > 0 THEN DO;
DO W = 1 TO BI;
WICKB = WICKB*(BI - W + 1);
END;
END;
IF BI = 0 THEN WICKB = 1;

*WICKC EQUALS THE FACTORIAL OF THE NUMBER NOT BIASED AGAINST (CI);
CI = NUMCONT - BI;
IF CI > 0 THEN DO;
DO W = 1 TO CI;
WICKC = WICKC*(CI - W + 1);
END;
END;
IF CI = 0 THEN WICKC = 1;

*‘NUMTRIAL’ EQUALS THE NUMBER OF WAYS OF HAVING WICKB ITEMS BIASED AGAINST
GIVEN NUMBIAS BIASED ITEMS;
NUMTRIAL = WICKA / (WICKB * WICKC);

*WE NEED TO DETERMINE ALL OF THE POSSIBLE COMBINATIONS OF HAVING ‘BI’ OF
‘NUMBIAS’ BIASED ITEMS. FOR EACH COMBINATION, THE ITEMS DEEMED TO BE BIASED
ARE INDICATED BY ‘IT’. SO, WE NEED TO DETERMINE ‘IT’, WHERE ‘IT’ TAKES ON ALL
INTEGER VALUES OF THE ITEMS BIASED AGAINST;
*LET'S START IT OFF BY SETTING 'IT' EQUAL TO THE ITEM NUMBER FOR THOSE ITEMS SELECTED IN THE COMBINATION, AND EQUAL TO 100 IF NOT;

DO K = 1 TO NUMCONT;
IT{K} = 100;
END;

DO K = 1 TO BI;
IT{K} = K;
END;

*NOW WE NEED TO FIND THE NEXT COMBINATION OF 'BI' BIASED ITEMS FROM A TOTAL OF 'NUMBIAS'. TO DO THIS, WE INCREMENT THE LAST IT{BI} BY ONE;

*LET'S FIRST DETERMINE WHERE THE INCREMENT MUST START FROM. WHAT FOLLOWS ACTS TO ASSIGN THE CONSTANT 'POSITION' THE VALUE OF WHICH ONE OF THE 'BI' BIASED ITEMS MUST SHIFT;

DO NT = 1 TO NUMTRIAL;

POSITION = BI;
DO W = 1 TO BI;
IF IT{W} > (NUMCONT - BI + W - 1) THEN POSITION = POSITION - 1;
ELSE POSITION = POSITION;
END;

IF NT = 1 THEN IT{BI} = IT{BI} - 1;
ELSE IT{BI} = IT{BI};

IF BI = NUMCONT THEN POSITION = BI;

IT{POSITION} = IT{POSITION} + 1;
ADD = 1;

DO K = 1 TO (POSITION - 1);
IT{K} = IT{K};
END;

IF POSITION < BI THEN DO:
DO K = (POSITION + 1) TO BI;
IT{K} = IT{POSITION} + ADD;
ADD = ADD + 1;
END;
END;

PRA1 = 1;

*SET UNDEFINED VALUES OF 'IT' TO 100;
DO J = 1 TO 12;
IF IT{J} = . THEN IT{J} = 100;
ELSE IT{J} = IT{J};
END:
*FIRST WE CONSIDER THE PROBABILITY OF MISSING FROM THE HIGHER VALID SCORE;

DO J = 1 TO (NUMCONT);

*ESTIMATE THE MANIFEST BIAS USING THE DIFFERENCE BETWEEN THE REFERENCE AND
FOCAL GROUP PROBABILITY OF CORRECT RESPONSE. HERE, THE MANIFEST BIAS IS
COMPUTED FOR EACH BIASED ITEM (J + 1) AT EACH VALID SCORE CATEGORY (I + BI);
BIAS = PRM{J+1, I+BI} - PFM{J+1, I+BI};

IF NRATM{I+BI} = 0 THEN BIAS = 1;
IF NFATM{I+BI} = 0 THEN BIAS = 1;

IF NRATM{I+BI} = 0 THEN BIAS = 1;
IF NFATM{I+BI} = 0 THEN BIAS = 1;

*SET J EQUAL TO ALL ITEMS BIASED AGAINST;
IF J = IT1 OR J = IT2 OR J = IT3 OR J = IT4 OR J = IT5 OR J = IT6
OR J = IT7 OR J = IT8 THEN PRA1 = PRA1*BIAS;
ELSE PRA1 = PRA1*(1 - BIAS);
END;

IF PRA1 = 0 THEN PRA1 = 0;
ELSE PRA1 = PRA1;

*COMPUTE THE TOTAL PROBABILITY 'PRB' OF BEING BIASED AGAINST ON 'BI' ITEMS GIVEN
YOU BELONG TO VALID SCORE CATEGORY 'I';
PRB{BI, I} = PRA1 + PRB{BI, I};

END: *THAT ENDS ONE TRIAL;

*HERE WE DETERMINE THE NUMBER OF FOCALS AT OBSERVED
SCORE M FROM THE HIGHER VALID SCORE CATEGORIES;

*COMPUTE THE NUMBER OF FOCALS AT OBSERVED SCORE M (TTN) WHO MISSED 'BI' ITEMS
DUE TO BIAS;
TTN{BI, I} = NFATM{I + BI} * PRB{BI, I};

IF TTN{BI, I} = 0 THEN TTN{BI, I} = 0;
ELSE TTN{BI, I} = TTN{BI, I};

*KEEP A RUNNING TALLY (PREB) OF THE NUMBER OF FOCAL GROUP MEMBERS AT
OBSERVED SCORE CATEGORY M WHO BELONG TO HIGHER VALID SCORE CATEGORIES;
PREB{I} = PREB{I} + TTN{BI, I};

IF PR{I, I+BI} = 0 THEN PR{I, I+BI} = 0;
COMPUTE THE NUMBER OF SUCCESSES ON THE TARGET ABILITY DIMENSION FOR FOCALS AT OBSERVED SCORE M AND VALID SCORE M + BI;
NHIGH(I) = NHIGH(I) + TTN{BI, I}*PR{1, I+BI};

END; *THAT ENDS CHOOSE BI;

ORIGIN EQUALS THE NUMBER OF FOCALS AT OBSERVED SCORE M WHO HAVE VALID SCORE ALSO EQUAL TO M;
ORIGIN(I) = NFATM(I) - PREB{I};

COMPUTE THE NUMBER OF SUCCESSES ON THE TARGET ABILITY DIMENSION AT EACH SCORE CATEGORY:
VNCOR(I) = NHIGH(I) + (ORIGIN{I})*PR{1, I};

COMPUTE THE PROBABILITY OF SUCCESS ON THE TARGET ABILITY DIMENSION AT EACH OBSERVED SCORE CATEGORY:
VPCOR{I} = VNCOR{I} / NFATM{I};

END; *THAT ENDS THE LOOP BY SCORE CATEGORY;

/*////////////////////////////////////////////////////////////////////////////////
*ESTIMATION OF THE CORRECTED REFERENCE NUMBERS;
*////////////////////////////////////////////////////////////////////////////////

DO I = 1 TO (NUMITEM + 1);

*COMPUTING THE ADJUSTED NUMBER OF CORRECT REFERENCE RESPONSES;
ACOR{I} = INT((VPCOR{I}*NRATM{I}) + .5);

IF ACOR{I} = . THEN ACOR{I} = 0;
ELSE ACOR{I} = ACOR{I};

*COMPUTING THE ADJUSTED NUMBER OF INCORRECT REFERENCE RESPONSES;
BCOR{I} = NRATM{I} - ACOR{I};

END;

/*////////////////////////////////////////////////////////////////////////////////
*CALCULATION OF THE CORRECTED MHD-DIF;
*////////////////////////////////////////////////////////////////////////////////

NUMCOR = 0;
DENOMCOR = 0;

DO J = 1 TO (NUMITEM + 1);
NUMCOR{J} = (ACOR{J} * DOBS{J}) / TOTOBS{J};

IF NUMCOR{J} = . THEN NUMCOR{J} = 0;
ELSE NUMCOR{J} = NUMCOR{J};

DENCOR{J} = (BCOR{J} * COBS{J}) / TOTOBS{J};
IF DENCOR{J} = . THEN DENCOR{J} = 0;
ELSE DENCOR{J} = DENCOR{J};

NUMERCOR = NUMERCOR + NUMCOR{J};
DENOMCOR = DENOMCOR + DENCOR{J};

END;

*THE CORRECTED MH ODDS RATIO (MHCOR);
MHCOR = NUMERCOR / DENOMCOR;

*THE CORRECTED LOG ODDS RATIO AND MHD-DIF (DDIFCOR);
LORCOR = LOG(MHCOR);
DDIFCOR = -2.35 * LORCOR;

*CALCULATION OF THE STANDARD ERROR FOR THE CORRECTED MANTEL HAENSZEL LOR;

DO J = 1 TO (NUMITEM + 1);
K1COR{J} = ((ACOR{J} * DOBS{J}) + (MHOBS * BCOR{J} * COBS{J}))/((TOTOBS{J}) ** 2);
K2COR{J} = (ACOR{J} + DOBS{J}) + (MHOBS * (BCOR{J} + COBS{J}));
K3COR{J} = K1COR{J} * K2COR{J};

IF K3COR{J} = . THEN K3COR{J} = 0;
ELSE K3COR{J} = K3COR{J};

K4COR = K4COR + K3COR{J};

END;

*COMPUTE THE STANDARD ERROR OF THE ADJUSTED MHD-DIF (SDDDIFCO);
SDLORCOR = SQRT((1/(2*NUMERCOR**2)) * K4COR);
SDDDIFCO = 2.35 * SDLORCOR;

*CORRECTED ETS CLASSIFICATION;

*TAKE THE ABSOLUTE VALUE OF THE CORRECTED MHD-DIF;
CABSDDIF = ABS(DDIFCOR);

*COMPUTE THE Z-SCORE FOR THE ADJUSTED MHD-DIF;
CZSCORE = ABS(DDIFCOR/SDDDIFCO);
CBSDF1 = ABS((CABSDDIF - 1)/SDDDIFCO);

*ASSIGN AN ETS CATEGORY, A = 1, B = 2, C = 3;
IF CABSDDIF < 1 OR CZSCORE <=2 THEN ETSCATCO = 1;
ELSE IF CABSDDIF >= 1.5 AND CBSDF1 > 2 THEN ETSCATCO = 3;
ELSE ETSCATCO = 2;

*KEEP A RUNNING TALLY OF THE NUMBER OF TRIALS IN EACH ETS CATEGORY. THE CATEGORIES A, B, AND C ARE DENOTED BY 'ETSCOA', 'ETSCOB', 'ETSCOC';
IF ETSCATCO = 1 THEN ETSCOA = 1;
ELSE ETSCOA = 0;

IF ETSCATCO = 2 THEN ETSCOB = 1;
ELSE ETSCOB = 0;

IF ETSCATCO = 3 THEN ETSCOC = 1;
ELSE ETSCOC = 0:

*//////////////////////////////////////////////////////////////////////////////////////////////;//
*CALCULATION OF THE CORRECTED MH CHI-SQUARE;
*/////////////////////////////////////////////////////////////////////////////////////////;//

DO J = 1 TO (NUMITEM + 1);
ABCOR{J} = ACOR{J} + BCOR{J};
ACCOR{J} = ACOR{J} + COBS{J};
CDCOR{J} = COBS{J} + DOBS{J};
BDCOR{J} = BCOR{J} + DOBS{J};
ALLCOR{J} = ACOR{J} + BCOR{J} + COBS{J} + DOBS{J};
EXPAC{J} = ((ACCOR{J}) * (ABCOR{J}))/ALLCOR{J};
VARAC{J} = ((ABCOR{J})*(ACCOR{J})*(CDCOR{J})*(BDCOR{J}))/((ALLCOR{J}**2)*(ALLCOR{J} - 1));
END:

TOTAC = SUM (OF ACOR1-ACOR4);
TVARAC = SUM (OF VARAC1-VARAC4);
TEXPAC = SUM (OF EXPAC1-EXPAC4);

*CALCULATION OF THE CORRECTED MH CHI-SQUARE;
MHCHICOR = ((ABS(TOTAC - TEXPAC) -.5)**2)/TVARAC;

*DETERMINING THE SIGNIFICANCE OF THE ADJUSTED MH CHI-SQUARE (SIGCOR);
IF MHCHICOR > 3.84 THEN SIGCOR = 1;
ELSE SIGCOR = 0;

OUTPUT:

END: *THAT ENDS THE LOOP FOR TRIALS;

*/////////////////////////////////////////////////////////////////////////////////////////;//
*COMPUTE MEANS AND SIGNIFICANCE RATES;
*/////////////////////////////////////////////////////////////////////////////////////////;//

PROC MEANS MEAN STD;
BY GRP;
VAR GRP DNS NUMCONT DDIFOBS DDIFCOR OZSCORE CZSCORE ETSOBA ETSOBB ETSOBC SIG ETSCOA ETSCOB ETSCOC SIGCOR;
RUN;
Appendix C

Simulation Program for Study Three

DATA SIM201B:

*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;
*CONSTANTS;
*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

NUMITEM = 40;
NUMEX = 500;
NUMGRP = NUMEX/2;
THETA = 1;
DS = .6;
NUMCONT = 8;
DNS = .6;

*NUMITEM IS THE NUMBER OF ITEMS ON THE TEST;
*NUMEX IS THE NUMBER OF EXAMINEES;
*NUMGRP IS THE NUMBER OF EXAMINEES IN EACH GROUP;
*THETA IS THE DIFFERENCE IN THE MEANS OF THE REFERENCE AND FOCAL GROUP ABILITY DISTRIBUTIONS;
*DS IS THE DIF INTRODUCED INTO THE STUDIED ITEM;
*DNS IS THE DIF INTRODUCED INTO THE NON-STUDIED ITEMS;
*NUMCONT IS THE NUMBER OF NON-STUDIED CONTAMINATING ITEMS;

*SET THE ETS CATEGORY COUNTERS TO ZERO;
ETSOBA = 0;
ETSOBB = 0;
ETSOBC = 0;

DO G = 1 TO 1000;

*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;
*LIST ALL ARRAYS;
*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~;

ARRAY TOTOB3{41} TOTOB3-TOTOB341;
ARRAY AOBS{41} AOBS1-AOBS41;
ARRAY BOBS{41} BOBS1-BOBS41;
ARRAY COBS{41} COBS1-COBS41;
ARRAY DOBS{41} DOBS1-DOBS41;

ARRAY NUMOBS{41} NUMOBS1-NUMOBS41;
ARRAY DEOBS{41} DEOBS1-DEOBS41;

ARRAY K1OBS{41} K1OBS1-K1OBS41;
ARRAY K2OBS{41} K2OBS1-K2OBS41;
ARRAY K3OBS{41} K3OBS1-K3OBS41;
ARRAY PROBB{41} PROB1-PROB41;
ARRAY OBS{41} OBS1-OBS41;
ARRAY Y{41} Y1-Y41;
ARRAY AA{41} AA1-AA41;
ARRAY APRE{41} APRE1-APRE41;
ARRAY BB{41} BB1-BB41;
ARRAY PFATM{41} PFATM1-PFATM41;
ARRAY NRATM{51} NRATM1-NRATM51;
ARRAY NFATM{51} NFATM1-NFATM51;
ARRAY DIA{41} DIA1-DIA41;
ARRAY D2A{41} D2A1-D2A41;

ARRAY PNMIS{41} PNMIS1-PNMIS41;
ARRAY VPCOR{41} VPCOR1-VPCOR41;
ARRAY VNCOR{41} VNCOR1-VNCOR41;
ARRAY NHIGH{41} NHIGH1-NHIGH41;
ARRAY NNMSI{41} NNMSI1-NNMSI41;

ARRAY TTPROB{9} TTPROB1-TTPROB9;
ARRAY DELTA{9} DELTA1-DELTA9;
ARRAY NUMTR{9} NUMTR1-NUMTR9;

ARRAY PR{9, 51} PR1-PR459;
ARRAY A{9, 41} A1-A369;
ARRAY B{9, 41} B1-B369;
ARRAY C{9, 41} C1-C369;
ARRAY D{9, 41} D1-D369;
ARRAY TTN{9, 41} TTN1-TTN369;
ARRAY TOT{9, 41} TOT1-TOT369;
ARRAY PRB{9, 41} PRB1-PRB369;

ARRAY IT{13} IT1-IT13;
ARRAY CAT{13} CAT1-CAT13;

ARRAY PREB{41} PREB1-PREB41;
ARRAY TTCOR{41} TTCOR1-TTCOR41;
ARRAY TTTPR{41} TTTPR1-TTTPR41;
ARRAY OKB{41} OKB1-OKB41;
ARRAY OKC{41} OKC1-OKC41;
ARRAY CORPR{41} CORPR1-CORPR41;
ARRAY ACOR{41} ACOR1-ACOR41;
ARRAY BCOR{41} BCOR1-BCOR41;
ARRAY NUMCOR{41} NUMCOR1-NUMCOR41;
ARRAY DENCOR{41} DENCOR1-DENCOR41;
ARRAY NFMU{81} NFMU1-NFMU81;

ARRAY K1COR{41} K1COR1-K1COR41;
ARRAY K2COR{41} K2COR1-K2COR41;
ARRAY K3COR{41} K3COR1-K3COR41;
ARRAY ORIGN{41} ORIGN1-ORIGN41;
ARRAY PRM(9, 51) PRM1-PRM459;
ARRAY PFM(9, 51) PFM1-PMF459;
ARRAY DIFM(9, 41) DIFM1-DIFM369;
ARRAY DELA(9, 41) DELA1-DELA369;
ARRAY DELB(9, 41) DELB1-DELB369;
ARRAY DELC(9, 41) DELC1-DELC369;
ARRAY SUMA(9) SUMA1-SUMA9;
ARRAY DELCH(9) DELCH1-DELCH9;
ARRAY DELCHA(9) DELCHA1-DELCHA9;
ARRAY SUMB(9) SUMB1-SUMB9;
ARRAY WACK(9) WACK1-WACK9;
ARRAY VACK(9) VACK1-VACK9;

ARRAY BLUB(41) BLUB1-BLUB41;
ARRAY WIPA(41) WIPA1-WIPA41;
ARRAY WIPB(41) WIPB1-WIPB41;

*////////////////////////////////////////////////////////////////////////////////////////////////////////
*SET ALL COUNTERS TO ZERO;
*////////////////////////////////////////////////////////////////////////////////////////////////////////

DO J = 1 TO 41;
TOTOBS{J} = 0;
AOBS{J} = 0;
BOBS{J} = 0;
COBS{J} = 0;
DOBS{J} = 0;
NUMOBS{J} = 0;
DENOBS{J} = 0;
END;

DO J = 1 TO 9;
TTPROB{J} = 0;
*D3NS{J} = 0;
*D4NS{J} = 0;
END;

DO J = 1 TO 9;
DO K = 1 TO 41;
PRB{J, K} = 0;
TOT{J, K} = 0;
A{J, K} = 0;
B{J, K} = 0;
C{J, K} = 0;
D{J, K} = 0;
END;
END;

*////////////////////////////////////////////////////////////////////////////////////////////////////////
*SIMULATE ITEM RESPONSES FOR EACH PERSON;
*////////////////////////////////////////////////////////////////////////////////////////////////////////

122
*SAMPLE THE DISCRIMINATION (AA) AND DIFFICULTY (BB) PARAMETERS;

*SAMPLE THE PARAMETERS FOR EACH ITEM;
DO L = 1 TO NUMITEM;

*SAMPLE THE DIFFICULTY PARAMETER FROM N(0, 1);
BB{L} = RANNOR(0);

*AA IS SAMPLED FROM A LOG-NORMAL DISTRIBUTION WHERE
AA = EXP(APRE). AND APRE=N(0,1225);
APRE{L} = RANNOR(0);
APRE{L} = APRE{L} *.1225);
AA{L} = EXP(APRE{L});

END:

*COMPUTE THE PROBABILITY OF CORRECT RESPONSE FOR EACH EXAMINEE;
DO N = 1 TO NUMEX;

*SAMPLE EACH EXAMINEE'S THETA VALUE (X) FROM N(0, 1);
X = RANNOR(0);

*SAMPLE A UNIFORM VARIATE FOR EACH ITEM;
DO L = 1 TO (NUMITEM + 1);
Y{L} = RANUNI(0);
END:

*ASSIGN GROUP MEMBERSHIP TO FOCAL (1) AND REFERENCE (2) MEMBERS;
IF N < (NUMGRP + 1) THEN GROUP = 1;
ELSE GROUP = 2;

*ADJUST THETA VALUES DEPENDING ON EQUALITY OF GROUP ABILITY DISTRIBUTIONS;
IF N < (NUMGRP + 1) THEN X = X - THETA;
ELSE X = X;

*PROBABILITY OF CORRECT RESPONSE FOR CONTAMINATED NON-STUDIED ITEMS;
DO I = 2 TO (NUMCONT+1);
IF GROUP = 1 THEN PROBB{I} = .2 + (.8)*(1/(1 + (EXP((-1.7)*(AA{I})*(X - BB{I} - DNS)))));
ELSE PROBB{I} = .2 + (.8)*(1/(1 + (EXP((-1.7)*(AA{I})*(X - BB{I})))));
END;

*PROBABILITY OF CORRECT RESPONSE FOR UNCONTAMINATED ITEMS;
DO I = (NUMCONT + 2) TO (NUMITEM);
PROBB{I} = .2 + (.8)*(1/(1 + (EXP((-1.7)*(AA{I})*(X - BB{I})))));
END;

*PROBABILITY OF CORRECT RESPONSE FOR STUDIED ITEM;
IF GROUP = 1 THEN PROBB{I} = .2 + (.8)*(1/(1 + (EXP((-1.7)*(AA{I})*(X - BB{I} - DS)))));
ELSE PROBB{I} = .2 + (.8)*(1/(1 + (EXP((-1.7)*(AA{I})*(X - BB{I})))));

*DETERMINE THE MANIFEST ITEM SCORE FOR EACH ITEM (OBS);
DO I = 1 TO NUMITEM;
IF Y(I) <= PROBB(I) THEN OBS(I) = 1;
ELSE OBS(I) = 0;
END;

*SET ALL UNDEFINED ITEM SCORES TO ZERO;
DO I = 1 TO NUMITEM;
IF OBS(I) = . THEN OBS(I) = 0;
ELSE OBS(I) = OBS(I);
END;

*CALCULATE TEST SCORES (OBSCORE);
OBSCORE = SUM (OF OBS1-OBS60);

*/////////////////////////////////////////////////////////////////////////////////
*COUNTERS FOR CELLS OF THE 2-BY-2 TABLES;
*STUDIED ITEM;
*/////////////////////////////////////////////////////////////////////////////////

*DETERMINE WHICH CELL OF THE TWO-BY-TWO TABLE THE EXAMINEE BELONGS IN. THE
TWO VARIABLES ARE GROUP (REFERENCE/FOCAL) AND ITEM RESPONSE
(CORRECT/INCORRECT). THE VARIABLE CODING CELL MEMBERSHIP (OBSCAT) HAS THE
FOLLOWING VALUES: REF/_COR = 1, REF/INCOR = 2, FOC/ COR = 3, FOC/INCOR = 4;
IF GROUP=2 AND OBS1=1 THEN OBSCAT=1;
ELSE IF GROUP=2 AND OBS1=0 THEN OBSCAT=2;
ELSE IF GROUP=1 AND OBS1=1 THEN OBSCAT=3;
ELSE OBSCAT=4;

*KEEP A ROLLING TALLY OF THE NUMBER OF OBSERVATIONS IN EACH CELL OF THE TWO-
BY-TWO TABLE. CONDITIONAL ON TEST SCORE. THE FOLLOWING ARRAYS ARE COUNTERS
FOR NUMBERS IN CELLS HAVING THE CODES 1, 2, 3, 4 ABOVE: 1 = AOBS, 2 = BOBS, 3 = COBS, 4
= DOBS. IN ADDITION, THE TOTAL NUMBER OF MEMBERS AT SCORE LEVEL 'J' IS STORED IN
'TOTOBJS';
J = OBSCORE + 1;
TOTOBJS(J) = TOTOBJS(J) + 1;
IF OBSCAT = 1 THEN AOBS(J) = AOBS(J) + 1;
ELSE IF OBSCAT = 2 THEN BOBS{J} = BOBS{J} + 1;
ELSE IF OBSCAT = 3 THEN COBS{J} = COBS{J} + 1;
ELSE DOBS{J} = DOBS{J} + 1;

*/////////////////////////////////////////////////////////////////////////////////
*COUNTERS FOR CELLS OF THE 2-BY-2 TABLES;
*CONTAMINATING ITEMS;
*/////////////////////////////////////////////////////////////////////////////////

*DO COUNTER FOR EACH CONTAMINATED ITEM;
DO K = 1 TO (NUMCONT+1);

*DETERMINE CELL MEMBERSHIP;
IF GROUP=2 AND OBS{K}=1 THEN CAT{K}=1;
ELSE IF GROUP=2 AND OBS{K}=0 THEN CAT{K}=2;
ELSE IF GROUP=1 AND OBS(K)=1 THEN CAT(K)=3;
ELSE CAT(K)=4;

*ROLLING COMPIlATION OF THE VALUES OF OBSERVED A B C D T;
J = OBSCORE + 1;
TOT(K, J) = TOT(K, J) + 1;
IF CAT(K) = 1 THEN A(K, J) = A(K, J) + 1;
ELSE IF CAT(K) = 2 THEN B(K, J) = B(K, J) + 1;
ELSE IF CAT(K) = 3 THEN C(K, J) = C(K, J) + 1;
ELSE D(K, J) = D(K, J) + 1;

END: *THAT ENDS THE LOOP FOR CONTAMINATING ITEMS;

END; *THAT ENDS THE LOOP FOR SUBJECTS;

*%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%;
*%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%;
**CALCULATION OF THE MANTEL-HAENSZEL ODDS RATIO;
**%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%;
**%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%;

*%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%;
*PROBABILITIES OF SUCCESS FOR NON-STUDIED ITEMS;
*%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%;

*NUMBERS IN FOCAL (NFATM) AND REFERENCE (NRATM) GROUPS;
DO I = 1 TO (NUMITEM + 1);
NRATM(I) = AOBS(I) + BOBS(I);
NFATM(I) = COBS(I) + DOBS(I);
END;

*PROBABILITY OF SUCCESS FOR NON-STUDIED ITEMS FOR EACH ITEM (K) AND
OBSERVED SCORE CATEGORY (I);
DO K = 1 TO (NUMCONT + 1);
DO I = 1 TO (NUMITEM + 1);
PR(K, I) = A(K, I)/(NRATM(I));
END;
END;

*SET ALL UNDEFINED PROBABILITIES OF CORRECT RESPONSE TO ZERO;
DO K = 1 TO (NUMCONT+1);
DO I = 1 TO (NUMITEM + 1);
IF NRATM(I) = 0 THEN PR(K, I) = 0;
ELSE PR(K, I) = PR(K, I);
END;
END;

*COMPUTE THE PROBABILITY OF CORRECT RESPONSE FOR THE REFERENCE (PRM)
AND FOCAL (PFM) GROUPS FOR EACH BIASED ITEM (K) AT EACH LEVEL OF TOTAL
TEST SCORE (I);
DO K = 1 TO (NUMCONT + 1);
DO I = 1 TO (NUMITEM + 1);
PRM{K, I} = A{K, I}/NRTM{I};
PFM{K, I} = C{K, I}/NRTM{I};
IF PRM{K, I} = . THEN PRM{K, I} = 0;
IF PFM{K, I} = . THEN PFM{K, I} = 0;
END;
END;

*COMPUTING THE VALID PROBABILITY OF SUCCESS ON THE TARGET ABILITY DIMENSION USING A GENERALIZED BINOMIAL PROCEDURE;

DO I = 1 TO (NUMITEM + 1);
PREB{I} = 0;
NHIGH{I} = 0;
DO BI = 1 TO NUMCONT;
WICKA = 1;
WICKB = 1;
WICKC = 1;

*FIRST, WE DETERMINE THE NUMBER OF COMBINATIONS OF MISSING B ITEMS OF "NUMCONT" POSSIBLE ITEMS;

*NOTE THAT BI = NUMBER OF ITEMS BIASED AGAINST,
NUMCOMB = NUMBER OF POSSIBLE COMBINATIONS
OF BEING BIASED AGAINST B ITEMS, NUMBIAS = NUMBER
OF BIASED NON-STUDIED ITEMS ON THE TEST;

*HERE, WICKA EQUALS THE NUMBIAS FACTORIAL;
DO W = 1 TO NUMCONT;
WICKA = WICKA * (NUMCONT - W + 1);
END;

*WICKB EQUALS THE FACTORIAL OF THE NUMBER BIASED AGAINST (B);
IF BI > 0 THEN DO;
DO W = 1 TO BI;
WICKB = WICKB*(BI - W + 1);
END;
END;
IF BI = 0 THEN WICKB = 1;

*WICKC EQUALS THE FACTORIAL OF THE NUMBER NOT BIASED AGAINST (C);
CI = NUMCONT - BI;
IF CI > 0 THEN DO;
DO W = 1 TO CI;
WICKC = WICKC*(CI - W + 1);
END;
END;
IF CI = 0 THEN WICKC = 1;

*NUMTRIAL EQUALS THE NUMBER OF WAYS OF HAVING WICKB ITEMS BIASED AGAINST
GIVEN NUMBIAS BIASED ITEMS;
NUMTRIAL = WICKA / (WICKB * WICKC);

*WE NEED TO DETERMINE ALL OF THE POSSIBLE COMBINATIONS OF HAVING
'BI' OF 'NUMBIAS' BIASED ITEMS. FOR EACH COMBINATION, THE ITEMS DEEMED TO BE
BIASED ARE INDICATED BY 'IT'. SO, WE NEED TO DETERMINE "IT", WHERE "IT" TAKES ON
ALL VALUES OF THE ITEMS BIASED AGAINST;

*LET'S START IT OFF BY SETTING 'IT' EQUAL TO THE ITEM NUMBER
FOR THOSE ITEMS SELECTED IN THE COMBINATION, AND EQUAL TO 100 IF NOT;

DO K = 1 TO NUMCONT;
IT[K] = 100;
END;

DO K = 1 TO BI;
IT[K] = K;
END;

*NOW WE NEED TO FIND THE NEXT COMBINATION OF 'BI' BIASED ITEMS FROM A TOTAL OF
NUMBIAS. TO DO THIS, WE INCREMENT THE LAST IT[BI] BY ONE;

*LET'S FIRST DETERMINE WHERE THE INCREMENT MUST START FROM. WHAT FOLLOWS
ACTS TO ASSIGN THE CONSTANT 'POSITION' THE VALUE OF WHICH ONE OF THE 'BI' BIASED
ITEMS MUST SHIFT;

DO NT = 1 TO NUMTRIAL;

POSITION = BI;
DO W = 1 TO BI;
IF IT[W] > (NUMCONT - BI + W - 1) THEN POSITION = POSITION - 1;
ELSE POSITION = POSITION;
END;

IF NT = 1 THEN IT[BI] = IT[BI] - 1;
ELSE IT[BI] = IT[BI];

IF BI = NUMCONT THEN POSITION = BI;

IT[POSITION] = IT[POSITION] + 1;
ADD = 1;

DO K = 1 TO (POSITION - 1);
IT[K] = IT[K];
END;

IF POSITION < BI THEN DO;
DO \( K = (\text{POSITION} + 1) \) TO \( \text{BI} \);
\( \text{IT}(K) = \text{IT}(\text{POSITION}) + \text{ADD} \);
\( \text{ADD} = \text{ADD} + 1 \);
END;
END;

\( \text{PRA}1 = 1 \);

DO \( J = 1 \) TO 12;
IF \( \text{IT}(J) = . \) THEN \( \text{IT}(J) = 100 \);
ELSE \( \text{IT}(J) = \text{IT}(J) \);
END;

*FIRST WE CONSIDER THE PROBABILITY OF MISSING FROM THE HIGHER VALID SCORE;

DO \( J = 1 \) TO (\( \text{NUMCONT} \));

*ESTIMATE MANIFEST BIAS USING THE DIFFERENCE BETWEEN THE REFERENCE AND
FOCAL GROUP PROBABILITY OF CORRECT RESPONSE;
\( \text{BIAS} = \text{PRM}(J+1, \text{I}+\text{BI}) - \text{PFM}(J+1, \text{I}+\text{BI}) \);  

IF \( \text{NRATM}(I+\text{BI}) = 0 \) THEN \( \text{BIAS} = 1 \);
IF \( \text{NFATM}(I+\text{BI}) = 0 \) THEN \( \text{BIAS} = 1 \);

IF \( \text{NRATM}(I+\text{BI}) = . \) THEN \( \text{BIAS} = 1 \);
IF \( \text{NFATM}(I+\text{BI}) = . \) THEN \( \text{BIAS} = 1 \);

*SET \( I \) EQUAL TO ALL ITEMS BIASED AGAINST:
IF \( J = \text{IT}1 \) OR \( J = \text{IT}2 \) OR \( J = \text{IT}3 \) OR \( J = \text{IT}4 \) OR \( J = \text{IT}5 \) OR \( J = \text{IT}6 \)
OR \( J = \text{IT}7 \) OR \( J = \text{IT}8 \) THEN \( \text{PRA}1 = \text{PRA}1*\text{BIAS} \);
ELSE \( \text{PRA}1 = \text{PRA}1*(1 - \text{BIAS}) \);
END;

IF \( \text{PRA}1 = . \) THEN \( \text{PRA}1 = 0 \);
ELSE \( \text{PRA}1 = \text{PRA}1 \);

*COMPUTE THE TOTAL PROBABILITY OF BEING BIASED AGAINST (\( \text{PRB} \)) ON ‘\( \text{BI}’ \) ITEMS;
\( \text{PRB}(\text{BI}, I) = \text{PRA}1 + \text{PRB}(\text{BI}, I) \);
END; *THAT ENDS ONE TRIAL;

*\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\n

*COMPUTE THE NUMBER OF FOCALS AT OBSERVED SCORE M WHO MISSED
‘\( \text{BI}’ \) ITEMS DUE TO BIAS;
\( \text{TTN}(\text{BI}, I) = \text{NFATM}(I + \text{BI})*\text{PRB}(\text{BI}, I) \);

IF \( \text{TTN}(\text{BI}, I) = . \) THEN \( \text{TTN}(\text{BI}, I) = 0 \);
ELSE \( \text{TTN}(\text{BI}, I) = \text{TTN}(\text{BI}, I) \);

128
*KEEP A ROLLING TALLY OF THE ESTIMATED NUMBER OF FOCALS AT OBSERVED SCORE M;
PREB\{I\} = PREB\{I\} + TTN\{BI, I\};

IF PR\{1, I-BI\} = . THEN PR\{1, I+B1\} = 0;

*COMPUTING THE NUMBER OF SUCCESSES ON THE TARGET ABILITY DIMENSION FOR FOCALS AT OBSERVED SCORE M AND VALID SCORE M + B;
NHIGH\{I\} = NHIGH\{I\} + TTN\{BI, I\}*PR\{1, I+B1\};

END; *THAT ENDS CHOOSE BI;

*ORIGN EQUALS THE NUMBER OF FOCALS AT OBSERVED SCORE M WHO WERE NOT BIASED AGAINST;
ORIGN\{I\} = NFATM\{I\} - PREB\{I\};

*COMPUTING THE NUMBER OF SUCCESS ON THE TARGET ABILITY DIMENSION AT EACH SCORE CATEGORY:
VNCOR\{I\} = NHIGH\{I\} + (ORIGN\{I\})*PR\{1, I\};

*COMPUTING THE PROBABILITY OF SUCCESS ON THE TARGET ABILITY DIMENSION AT EACH OBSERVED SCORE CATEGORY:
VPCOR\{I\} = VNCOR\{I\} / NFATM\{I\};

END; *THAT ENDS THE LOOP BY SCORE CATEGORY;

*-----------------------------------;
*CALCULATION OF MB-DIF;
*-----------------------------------;

WIPAA = 0;
BLUBN = 0;

DO K = 1 TO (NUMITEM+1);

*COMPUTE THE VALID PROBABILITY OF CORRECT RESPONSE WHERE

WIPA IS THE WEIGHT ASSIGNED TO EACH TEST SCORE CATEGORY;
WIPA\{K\} = ((NFATM\{K\})*NRATM\{K\})/(NRATM\{K\}+NFATM\{K\}));
IF NRATM\{K\} = . OR NFATM\{K\} = . THEN WIPA\{K\} = 0;

*COMPUTING THE WEIGHTED MANIFEST BIAS AT SCORE CATEGORY K;
BLUB\{K\} = (WIPA\{K\})*(VPCOR\{K\} - PFM\{1, K\});
IF VPCOR\{K\} = . OR PFM\{1, K\} = . THEN BLUB\{K\} = 0;

*SUMMING THE WEIGHTED MANIFEST BIASES;
BLUBN = BLUBN + BLUB\{K\};

*SUMMING THE WEIGHTS ACROSS SCORE CATEGORIES;
IF WIPA\{K\} = . THEN WIPA\{K\} = 0;
WIPAA = WIPAA + WIPA\{K\};
*COMPUTING MB-DIF:
MBDIF = BLUBN/WIPAA;

*STANDARD ERROR FOR MB-DIF;

WIPAA = 0;
WIPBB = 0;
NNNF = 0;

DO K = 1 TO (NUMITEM+1);

*COMPUTING THE WEIGHT FOR EACH SCORE CATEGORY:
WIPA[K] = ((NFATM{K}*NRATM{K})/(NRATM{K}+NFATM{K}));
IF NRATM{K} = . OR NFATM{K} = . THEN WIPA{K} = 0;

WIPB{K} = (WIPA{K}**2)*(((PFM{1, K}*(1-PFM{1, K}))/NFATM{K}) + ((VPCOR{K}*(1-VPCOR{K}))/NRATM{K}));
IF WIPA{K} = . THEN WIPA{K} = 0;
IF WIPB{K} = . THEN WIPB{K} = 0;

WIPAA = WIPAA + WIPA{K};
WIPBB = WIPBB + WIPB{K};

END;

*COMPUTING THE STANDARD ERROR OF MB-DIF (SDMB);
SDMB = ((WIPBB)/(WIPAA**2))**.5;

*DETERMINING THE SIGNIFICANCE OF MB-DIF (SIG);
IF MBDIF > 1.96*(SDMB) THEN SIG = 1;
ELSE SIG = 0;

OUTPUT:
END:

*COMPUTE MEANS AND PROPORTION OF SIGNIFICANT ITEMS;

PROC MEANS MEAN STD;
BY GRP;
VAR DNS NUMCONT MBDIF SDMB SIG;
RUN;
Appendix D

Proof of the Relationship Between MB-DIF and STNDP-DIF

Manifest bias for the focal group at observed score $m = j$ is expressed as

$$(\delta m = j) = E(\Theta | G = F, m = j) - E(Y | G = F, m = j). \quad (D.1)$$

From Result 4.2 it is asserted that

$$E(\Theta | G = R, \nu = j) = E(Y | G = R, m = j). \quad (D.2)$$

Thus, only when $\nu = m$ for the focal group will it be true that

$$(\delta m = j) = E(Y | G = R, m = j) - E(Y | G = F, m = j). \quad (D.3)$$

Since $\nu = m$ defines the case of no contamination, it follows that manifest bias is equal to the difference between the item test regressions for the focal and reference populations only when there is zero contamination.
Appendix E

Estimation Procedure for the Number of Focal Group Members at Valid Score Category $v = j$

The observed number of focal group examinees at observed score category $m = j$ can be expressed as

$$ (N|m = j) = \sum_{h=0}^{k} P(B|v = j + h, h)(N|v = j + h). \quad (E.1) $$

where all terms are conditional to the focal group, and $B$ is the event of incorrectly responding to $h$ of the $k$ biased items due to sub-threshold levels of expressed nuisance determinant. Using Equation E.1, the number of examinees at valid score category $v = j + h$ can be expressed as

$$ (N|v = j + h) = \frac{(N|m = j) - \sum_{l=0, l\neq h}^{k} P(B|v = j + l, l)(N|v = j + l)}{P(B|v = j + h, h)} \quad (E.2) $$

Since the number of focal group members at valid score category $v = j + h$ must be known to estimate each lower score category, Equation E.2 must be solved sequentially, solving first for the highest possible valid score category $(v + k)$, and then sequentially each lower score category.