A Modified Higher-Order DINA model for Detecting Differential Item Functioning and Differential Attribute Functioning

by

Feiming Li

(Under the direction of Allan S. Cohen)

Abstract

This dissertation presents a modified higher-order DINA model for separating the source of construct-relevant (i.e., benign) differential item functioning (DIF) from construct irrelevant (i.e., adverse) DIF. The model-based method provides a natural framework for detecting both differential attribute functioning (DAF) and DIF in a cognitive diagnostic modeling framework: DIF detection ensures test fairness and improves test validity in terms of group difference in item performance after conditioning on attribute mastery profiles, whereas DAF detection provides a good understanding of group strength and weakness in terms of a set of cognitive attributes after conditioning on general ability. An MCMC algorithm employing Gibbs sampling was used to estimate the new model, and a simulation study was done to examine model recovery, Type I error rates, and power under different testing conditions. For DIF detection, the model-based method was also compared with the MH method using two types of matching criteria, a total score as the matching criterion and an attribute profile as the matching criterion. Finally, a statewide mathematics test was used to illustrate the implementation and possible limitations of the new method.

Index words: Cognitive diagnostic assessment, modified higher-order DINA model, differential item functioning, differential attribute functioning

A Modified Higher-Order DINA model for Detecting Differential Item Functioning and Differential Attribute Functioning

by

Feiming Li

B.A., Zhejiang University, 1999

M.Ed., Zhejiang University, 2002

M.S., University of Georgia, 2005

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2008

© 2008

Feiming Li

All Rights Reserved

A Modified Higher-Order DINA model for Detecting Differential Item Functioning and Differential Attribute Functioning

by

Feiming Li

Approved:	
Major Professor:	Allan S. Cohen
Committee:	Deborah Bandalos Seock-Ho Kim Nicole Lazar Jonathan Templin

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2008

Dedication

To my lovely daughter, Qianqian.

Acknowledgments

I would like to express the deepest appreciation to my advisor, Dr. Allan Cohen. Without his persistent encouragement and support this dissertation would not have been possible. During my graduate study, he made endless efforts in guiding me to become a successful scholar: supporting me for many national and international conferences, spending his time in reading and revising the draft of my papers, helping me get my first paper accepted, and providing me with sincere opinions about my future career. He's even happier than me for any progress I have made. He continually and convincingly conveyed a spirit of adventure in regard to research. His humor and understanding made it possible for me to work in a happy and relaxed atmosphere.

I would like to thank my committee members, Dr. Seock-Ho Kim and Dr. Deborah Bandalos. I got most of training in REMS program from their courses, which provided me a solid foundation in educational measurement. I will not forget Dr. Kim's patience and his interesting philosophy about research and life. I also cannot forget Dr. Bandalos' warmth and encouragement. I was amazed by her sensitivity to an international student like me in understanding my questions and my points of view immediately.

I would like to thank my committee member, Dr. Jonathan Templin. He's been my guide to developing my interest in cognitive diagnostic assessment. He always generously shared his thoughts and his latest papers. These nurtured the ideas that grew into this dissertation.

I also would like to thank my colleagues in REMS program for their help and care in my study and personal life. Without them, I couldn't have been so happy during these years.

Finally I would like to thank my parents and my husband, who have always supported me and my decisions, allowing me to grow into the person I am today.

Table of Contents

List of Figures

List of Tables

Chapter 1

The Statement of Problem

In the past decade, a number of cognitive diagnostic models (CDMs) have been developed to evaluate examinees' status relative to mastery or non-mastery of a set of cognitive attributes, knowledge or skills (Dibello et al., 1995; Junker & Sijtsma, 2001; Hartz, 2002; de la Torre & Douglas, 2004; Templin, 2004; Henson et al., 2007). These models differ from item response theory (IRT) models, which linearly or partially order examinees in a low-dimensional latent space. Rather, CDMs provide more fine-grained information regarding individual or population-level learning weaknesses and strengths than is available in the standard unidimensional IRT models. As a result of this increase in information, the development of CDMs has provided a new perspective for study of some common problems in educational measurement. In this dissertation, we examine one of those problems, detection of differential item functioning (DIF).

Definition of DIF. DIF is defined as a difference in the probability of a correct response in one group compared to another group for examinees of the same ability (e.g., Pine, 1977). More generally, DIF is said to occur when the probability of a particular response is higher for one group relative to another group or groups, conditioned on ability (Chang, Roussos, & Mazzeo, 1996; Cohen, Kim, & Baker, 1993). Defined in this way, DIF is of interest because it is directly related to test fairness and validity.

Currently, there exist a number of non-parametric and parametric DIF detection procedures. The Mantel-Haenszel (MH), as modified by Holland and Thayer (1988), and SIBTEST (Shealy & Stout, 1993) are both appropriate non-parametric DIF detection procedures, for example, for use with tests that are based on classical test theory scoring models. Several

IRT-based procedures are available as well, including Lord's chi-square method (Lord, 1980), Raju's (Raju, 1988, 1990) area measures, and the likelihood ratio test (Thissen, Steinberg, & Wainer, 1988). Even though these methods have been found to be very useful for detection of DIF, little progress has been made in understanding the causes of DIF (Roussos & Stout, 1996). One objective in this dissertation is to determine whether CDMs may be useful in helping to understand the causes of DIF and what may be learned about DIF by their use.

Multidimensional Framework for DIF. Shealy and Stout (1993) provide a multidimensional framework in which DIF is said to occur mainly because the item measures dimensions that are secondary to a primary or target ability. This set of secondary dimensions has further been categorized into auxiliary and nuisance dimensions (Roussos & Stout, 1996; Douglas, Roussos, & Stout, 1996). The term auxiliary dimensions refers to dimension(s) that the test was intended to measure. The term nuisance dimensionality is used to refer to dimensions that the test was not intended to measure. These dimensions are not normally included in the test specifications or the measurement model and, as a result, are considered to be construct-irrelevant.

Shealy and Stout (1993) provide a mathematical definition of DIF as being present if both of the following two conditions are satisfied :

- The item is sensitive not only to the primary construct θ, but also to some secondary construct η, and
- A difference exists between groups of interest in their conditional distribution on η given a fixed value of θ (i.e., $\eta|\theta$).

According to this definition, θ is measured by the primary dimension, and DIF is caused by the η dimensions. In order to be the cause(s) of benign DIF, η would need to be auxiliary rather than nuisance dimensions.

Shealy and Stout (1993) referred to DIF that is caused by auxiliary dimensions as benign DIF, and to DIF that is caused by nuisance dimensions as adverse DIF. That is, benign

DIF serves to differentiate groups based on their difference on construct-related dimensions conditional on ability. Adverse DIF serves to differentiate groups based on differences on construct-irrelevant dimensions. In this way, adverse DIF contributes to lowering construct validity because irrelevant constructs are measured by the item. Benign DIF is viewed as enhancing construct validity of the test, only if the dimensionality that caused the benign DIF is modelled by the statistical model used to scale the test or reported as test performance (Walker & Beretvas, 2001). Benign DIF can be eliminated by conducting an additional DIF analysis in which all construct-relevant (i.e., auxiliary) dimensions including the primary dimension are modeled and included in the conditioning variable. In addition, construct validity can be enhanced by including these auxiliary dimensions in the reported scores. Adverse DIF, however, can be eliminated only by deleting the item or by revising it sufficiently to remove the unwanted dimensionality from being measured. It is important, furthermore, that the adverse DIF item not be included in the conditioning variable used for DIF analysis.

Motivation for this Study. This study has the following motivation: (1) Little research has focused on investigating the composition of factors of benign DIF. This is potentially an important area of study and could conceivable lead to methods for improving construct validity (Walker & Beretvas, 2001). For example, if two groups with known differences on some construct-relevant dimensions do show DIF on the items measuring those dimensions, it suggests that the items are capable of differentiating between these two groups on those dimensions. What would be useful, would be finding a way to isolate those factors which are the possible causes of benign DIF so that these factors might be included in the measurement model. In this way, test validity potentially could be improved. The detection of benign DIF also helps us understand the cause(s) of group differences in item performance in addition to that due to differences in the primary ability. Gierl et al. (2003), for example, applied a DIF analysis framework to identify gender differences in cognitive skills measured on a test of mathematics.

(2) Benign DIF is usually confounded with adverse DIF and, as a result, more items are detected as functioning differentially and removed or revised than might be necessary. Following the logic of Shealy and Stout (1993), only adverse DIF is harmful, and only adverse DIF items need to be removed from the test. Douglas et al. (1996) identified two requirements for determining if a DIF item has adverse DIF:

- The matching criterion does not result in a construct-valid matching of examinees on the construct intended to be measured by the test.
- The secondary dimension of the DIF item is a true nuisance dimension.

Most current DIF procedures consider only a unidimensional conditioning variable for comparing item performance among different groups. That is, either the observed total score or the latent ability on which item performance is conditioned are treated with a unidimensional measurement model, although many cognitive tests are to some extent multidimensional. In addition, previous research on DIF has indicated that a primary source of DIF is that due to the failure of the matching criterion to account for the complete latent ability space (Ackerman, 1992; Ackerman & Evans, 1994). When the matching variables don't represent the complete latent construct the test is intended to measure, and groups differ on the secondary dimensions conditional on the primary dimension, items that measure the auxiliary dimension(s), that is, items that are not measuring nuisance dimensions, will be flagged as DIF items. As a result, unnecessary cost may be incurred for re-editing or replacing those items.

3) Only two DIF detection studies have been reported using a cognitive diagnostic modeling framework. These results (discussed below) suggest important advantages from using this framework. Milewski and Baron (2002) investigated DIF from a CDM framework, extending DIF detection for detection of differential skill functioning (DSF). DSF was designed to examine group differences in skill mastery performance conditioned on overall ability rather than examining group differences in item response performance conditioned on overall ability. In the Milewski and Baron study, skill mastery profiles were estimated using a

CDM, and then DIF detection was done based on comparison on each skill between different manifest groups (e.g., schools, states) and the total population. In this way, the focal group was composed of an aggregated group of interest, such as students in the same school, and the reference group was composed of a random sample of the population. Though the term differential skill functioning was used to refer to cognitive skills, the intent of the Milewski and Baron study was not to explore whether a skill was biased, but rather to investigate strengths and weaknesses of different manifest groups. From this perspective, DSF was conceived of as a cause of benign DIF. Zhang (2007) extended this approach by employing attribute profiles estimated from a DINA model (Macready & Dayton, 1977; Haertel, 1989; Junker & Sijtsma, 2001) as matching variables for both the MH and SIBTEST procedures. By using the attribute profiles instead of a unidimensional ability as matching criteria, Zhang (2007) established a complete construct-valid latent space to detect only adverse DIF. Zhang's method showed reduced Type I error rates compared to error rates realized using total raw score as the matching criteria.

Simultaneous Detection of Adverse and Benign DIF. The utility of the CDM framework as a means of parsing the construct relevant and construct irrelevant variance measured by individual test items would appear to be potentially useful. The focus of this study, therefore, was to develop a method to simultaneously detect both DSF, the cause of benign DIF, and adverse DIF, in the context of a CDM. To do this, a higher-order Deterministic Inputs, Noisy And Gate (DINA) model (de la Torre & Douglas, 2004) was employed. de la Torre and Douglas (2004) extended the DINA model to include an IRT model as the higher-order model (i.e., a model inserted above the DINA model). This model has two levels in other words: The lower level model is a DINA model and the higher level model is a two-parameter logistic (2PL) IRT model. The DINA model is used to relate item responses to an attribute mastery profile. The 2PL is used to relate mastery status for each attribute to one or more general ability(-ies). An important assumption in this model is that item responses are independent, conditioned on the attribute mastery profile. Mastery of attributes, however, is assumed to be independent conditional on the general abilities.

This hierarchical relationship among items, attributes, and general ability provides a natural framework within which to detect both DIF and differential attribute functioning (DAF). The term DAF is used in this paper rather than DSF, replacing skill with the term attribute. This is because attribute is a more generic term for psychological construct, and refers to tasks, subtasks, skills or cognitive processes required by the test (Tatsuoka, 1995). DAF is defined as the differential probability of mastery of an attribute among groups of interest matched on general ability(-ies).

As noted above DIF has been defined as the probability of a correct response between two groups conditional on ability (e.g., Pine, 1977). In this study, we modify this definition to include the profile of cognitive attributes as the conditioning variables. It is assumed that the specification of attributes required for the test is correct, so that the profile of cognitive attributes can be regarded as a complete latent ability space. In this way, the use of this profile as conditioning variables excludes benign DIF and detects only adverse DIF. In the following, DIF, as studied by the new method developed in this study, consistently refers to the adverse DIF. A revised definition of DIF, therefore, is given by the following:

- DIF is defined as a differential probability for a particular response endorsed by one group compared to that endorsed by another group (or other groups) conditioned on the mastery status (or attribute) profile.

This definition does not confine DIF to occur only between two groups or for responses to be scored as correct or incorrect, and the responses can be being dichotomous, polytomous or even continuous.

DAF is defined as a differential propensity of one group to have a greater probability of mastery on an attribute compared to another group, conditioned on general ability. From this perspective, DIF is viewed as adverse DIF, since the group is conditioned on a measure of the complete construct-related dimensions (as reflected in the attribute mastery profiles). DAF is considered as the cause of benign DIF, since it is the difference in cognitive attributes conditional on the primary ability, and these cognitive attributes are construct-related dimensions intended to be measured by the test.

The higher-order DINA model operates in such a way that the item response function at the lower level is determined by the item parameters and the attribute mastery status at the higher level is determined by the attribute parameters. The techniques used to detect DIF and DAF in this study follow the spirit of item parameter equality comparisons used in Lord's Chi-square method (Lord, 1980). DIF and DAF were directly modeled by adjusting the higher-order DINA model rather than by calibrating the model separately for each group and then calculating differences.

Estimation of the Model. A Markov chain Monte Carlo (MCMC) algorithm employing Gibbs sampling was implemented to estimate all parameters in the model. One advantage of this method for this study is that empirical posterior distributions for all parameters, including estimates of DAF and DIF, were obtained from the post burn-in iterations. This differs from Lord's chi-square in that estimates of DIF are obtained based on statistics from an asymptotic chi-square distribution. The empirical posterior distributions for the parameters accounting for the differences can be directly used to test the significance of the DIF and the DAF.

Summary. In this research, (1) a new method was developed to separate the sources of benign DIF and adverse DIF, enabling simultaneous detection of DIF and DAF, (2) Type I error and power analyses were presented for the new method across conditions commonly found in typical testing situations, and (3) a real-data example was presented and interpreted in light of the simulation study and model assumptions.

The modified higher-order DINA model used in this study provides a natural framework within which to simultaneously separate the source of construct-relevant (i.e., benign) DIF from construct-irrelevant (i.e., adverse) DIF. DIF detection helps ensure test fairness and helps improve test validity, whereas DAF detection provides a better understanding of differential knowledge structures across groups. More specifically, this new method can be used to define group strengths and weaknesses in terms of a set of cognitive attributes after conditioning on general ability.

Chapter 2

Literature Review

2.1 Cognitive Diagnostic Models

Several cognitive diagnostic models (CDMs) have been developed to evaluate examinees' status relative to mastery or non-mastery on each of a set of attributes (Haertel, 1989; DiBello et al., 1995; de la Torre & Douglas, 2004; Hartz, 2002; Henson et al., 2007; Junker & Sijtsma, 2001; Templin, 2004). CDMs differ from IRT in that IRT models linearly or partially order examinees in a low-dimension latent space. CDMs, on the other hand, try to estimate a set of cognitive attributes which each examinee has mastered or not mastered based on the examinee's responses to the test items.

Some Terminology. Prior to introducing the cognitive diagnostic models, some key notation and terminology need to be clarified. An important characteristic of CDMs is their capability for providing a profile of cognitive attributes an examinee has mastered or not mastered. As noted in Chapter 1, the term attribute is used generically to refer tasks, subtasks, cognitive processes, or skills that are intended to be measured (Tatsuoka, 1995). The relationship between items and attributes is specified in the Q-matrix, first introduced by Tatsuoka (1990). The Q-matrix is a format for specifying the underlying cognitive attributes measured by the test items. If we have $i = 1, \ldots, I$ items and $k = 1, \ldots, K$ attributes, the Q-matrix is specified by an $I \times K$ matrix with q_{ik} as elements: $q_{ik} = 1$, when attribute k is required to respond correctly to item i and $q_{ik} = 0$ otherwise. Before a CDM is fit to test data, the Q-matrix must be determined. Typically, this is done by analysis of the attributes required for each item, based on the expert judgments, such as from content specialists.

In the literature review presented in this section, the DINA and the higher-order DINA models are described. As specified here, these models are only intended to be used for dichotomous item response data.

2.1.1 DINA Model

Junker and Sijtsma (2001) introduced the name DINA for a CDM that can be used to describe the probability of a correct response as a function of an examinees' attribute profiles and the item parameters. (See also Macready and Dayton (1977), Haertel (1989), and Doignon and Falmagne (1999) for earlier discussions of this model.) Applications of the DINA model along with MCMC algorithms for estimation of model parameters are given in Junker and Sijtsma (2001).

The inputs to the DINA model in the set of latent responses, ξ_{ij}'s, are each determined by 1) the elements of the Q-matrix, q_{ik}, which take values of 1 or 0, indicating whether attribute k is required or not required, respectively, and 2) a mastery or non-mastery status, $\alpha_{jk} = 1$ or 0, indicating whether or not examinee j has mastered attribute k. This model can be expressed as

$$\xi_{ij} = \prod_{k=1}^{K} \alpha_{jk}^{q_{ik}} \ , \tag{2.1}$$

where the deterministic latent response ξ_{ij} denotes whether examinee j has mastered all required attributes for item j. Thus, each item divides the population into two classes, those who master all required attributes for that item, indicated by $\xi_{ij} = 1$, and those who miss at least one required attribute for that item, indicated by $\xi_{ij} = 0$. The model is considered as stochastic, since the observed response Y_{ij} is not completely consistent with the latent response ξ_{ij}. The probabilistic relation is governed by two "Noisy" parameters unique to each item, s_i, a slip parameter, and g_i, a guessing parameter. As described for the signal detection model (Green & Swets, 1966), the ξ_{ij} are estimated from noisy observations, the Y_{ij}, by two error probabilities, a "slipping" parameter, s_i, indicating the probability of a false negative for item i, and a "guessing" parameter, g_i, indicating the false positive rate for item i. That is, s_i is the probability of missing item i for someone who is classified as mastering all required attributes (i.e., $\xi_{ij} = 1$):

$$s_i = P(Y_{ij} = 0|\xi_{ij} = 1) \ , \tag{2.2}$$

and g_i is the probability of a correct response for someone classified as lacking at least one required attribute (i.e., $\xi_{ij} = 0$):

$$g_i = P(Y_{ij} = 1|\xi_{ij} = 0) \ . \tag{2.3}$$

Given examinee parameter ξ_{ij} and item parameters s_i and g_i, the probability of a correct response to item i can be written as

$$P(Y_{ij} = 1|\xi_{ij}) = (1 - s_i)^{\xi_{ij}} g_i^{(1-\xi_{ij})} \tag{2.4}$$

That is, the probability of a correct response to item i can only be divided into two categories: g_i for any examinee j who misses one or more attributes required by item i (i.e., $\xi_{ij} = 0$), and $1 - s_i$ for any examinee j who masters all attributes required for item i (i.e., $\xi_{ij} = 1$). The model is conjunctive since the probability of a correct response is the same whether one or more than one of the required attributes is missed. In addition, mastering more attributes than required for correctly answering item i doesn't make the correct response probability higher either, so the model is noncompensatory.

2.1.2 Higher-Order DINA Model

de le Torre and Douglas (2004) extended the DINA model to include an IRT model for the joint distribution of the attributes. The result is a higher-order DINA model that assumes the several cognitive attributes are dependent on one or at most a small number of general abilities. The lower-order part of the model is the same as the DINA model, and the higher-order part is the same as the 2PL model given as

$$\text{Logit}[P(\alpha_{jk}|\theta_j)] = a_k\theta_j - \beta_k \ , \tag{2.5}$$

where α_{jk} is the indicator of mastery status for examinee j on attribute k, θ_j is the ability parameter for examinee j, and a_k and β_k are the discrimination and difficulty parameters, respectively, of attribute k. The higher-order DINA model is used to classify examinees on specific attributes and simultaneously provide estimates of general ability(-ies). In the original description by de la Torre and Douglas, the broadly-defined latent traits were assumed to consist of a small number of dimensions. Here we assume a unidimensional ability, θ, and attributes, α, to be independent conditional on θ.

2.2 Differential Item Functioning

Differential item functioning (DIF) has received considerable attention in the psychometric literature in large part because it is directly related to test fairness and validity. An item shows DIF if individuals from the reference and focal group have different probabilities of getting it correct conditioned on ability (Pine, 1977). The usual DIF analysis is designed to compare two groups, the focal group and the reference group. The focal group is generally the group that is the focus of concern, and the reference group serves as the group against which the focal group is compared. Methods for detecting DIF include non-parametric methods, such as the Mantel-Haenszel procedure (MH: Holland & Thayer, 1988) and the simultaneous item bias test (SIBTEST) procedure (Shealy & Stout, 1993), parametric non-IRT methods, including logistic regression (Swaminathan & Rogers, 1990) and parametric methods in the context of IRT such as the area measures by Raju (1988, 1990), and comparison of item parameters (Lord, 1980; Thissen, Steinberg, & Wainer, 1988). In this dissertation, only the MH method was used in conjunction with the new method based on the higher-order DINA model. In part, this is because different tests of DIF tend to differ slightly, even in the context of simulation studies, making it difficult to determine which solution is more accurate. In addition, the MH is often selected, in part, because it is easy to use and also because it has been found to work even in relatively small samples.

2.2.1 Mantel-Haenszel Method

The Mantel-Haenszel (MH) procedure was originally introduced by Mantel and Haenszel (1959) for the study of matched groups in the context of clinical cancer trials. Holland (1985) and Holland and Thayer (1988) adapted this procedure for assessing differential item functioning on tests. In the MH procedure, two contrasting examinee groups are identified, the reference group, r, and the focal group, f. The two groups are matched on some criterion that is assumed to accurately represent the construct of interest. In the usual application of the MH, the matching variable is the raw score (i.e., the total score) on the test and includes the studied item. The matching variable is used to divide the range of scores into K score groups for use in comparing the correct versus incorrect performance of the r and f groups. To do this, K 2×2 contingency tables are constructed. The MH chi-square is then computed as a single degree of freedom chi-square over the K 2×2 contingency tables. A 2×2 contingency table for score group k $(k = 1, \ldots, K)$ for the studied item is shown in Table 2.1.

Table 2.1: Contingency Table For Mantel-Haenszel DIF Statistic

Group	Right	Wrong	Total
Reference Group	R_{rk}	W_{rk}	N_{rk}
Focal Group	R_{fk}	W_{fk}	N_{fk}
Total Group	R_{tk}	W_{tk}	N_{tk}

In Table 2.1, R_{rk} and W_{rk} are counts of right and wrong responses, respectively, in the reference group at score level k, R_{fk} and W_{fk} are counts of right and wrong responses, respectively, in the focal group at score level k, R_{tk} and W_{tk} are the count of right and wrong responses in the total group at score level k. In addition, N_{rk}, N_{fk}, and N_{tk} are the number of examinees in the reference group, the focal group and the total group, respectively. The null hypothesis of the MH method for DIF analysis can be expressed as

$$H_0 : \frac{R_{rk}}{W_{rk}} = \frac{R_{fk}}{W_{fk}} \ . \tag{2.6}$$

That is, the odds of getting the item correct in the focal group is the same as that in the reference group at a given level of the matching variable. The MH method provides both a test of statistical significance and an estimate of effect size for DIF. For the test statistic, the MH method computes a chi-square statistic which is given by

$$MH_{\chi^2} = \frac{[|\sum_{k=1}^{K}(R_{rk} - E(R_{rk}))| - .5]^2}{\sum_{k=1}^{K} Var(R_{rk})} \ , \tag{2.7}$$

where

$$E(R_{rk}) = N_{rk}R_{rk}/N_{tk}, Var(R_{rk}) = \frac{N_{rk}N_{fk}R_{tk}W_{tk}}{N_{tk}^2(N_{tk} - 1)} \ . \tag{2.8}$$

Under the null hypothesis of no DIF, the statistic MH_{χ^2} has approximately a chi-square distribution with one degree of freedom. A significant MH_{χ^2} indicates uniform DIF, that is a difference in the probability of a correct answer to an item between two groups that is constant across all ability levels. A measure of effect size is also provided for the MH method in which the common odds ratio α_{MH} is calculated, representing the ratio of the odds that a member of the reference group will answer the studied item correctly to the odds that a matched member of the focal group will do the same. If $\alpha_{MH} = 1$, then there is no difference in the performance of the two groups on the item at the k^{th} score level. This value is easily combined across all score levels to obtain a measure of the DIF effect size, α_{MH}. The form of this effect size is given by the following formula:

$$\alpha_{MH} = \frac{\sum_{k=1}^{K} R_{rk}W_{fk}/N_{tk}}{\sum_{k=1}^{K} W_{rk}R_{fk}/N_{tk}} \ . \tag{2.9}$$

Holland and Thayer (1988) also proposed the following natural logarithmic transformation of the odds ratio α_{MH} to make the effect size scale symmetric:

$$\Delta\alpha_{MH} = -2.35 \ \ln(\alpha_{MH}) \ . \tag{2.10}$$

A value of zero indicates no DIF, a positive value indicates that the item favors the focal group, and a negative value indicates the item favors the reference group. Educational Testing Service (ETS) classifies DIF based on the $\Delta\alpha_{MH}$ into 3 levels (Dorans & Holland, 1993):

A. Negligible DIF, when chi square is not significant and $\mid \Delta\alpha_{MH} \mid < 1$.

B. Intermediate DIF, when chi-square is significant and $1 <\mid \Delta\alpha_{MH} \mid < 1.5$.

C. Large DIF, when chi-square is significant and $\mid \Delta\alpha_{MH} \mid \geq 1.5$.

The ETS standards were used in the real data example presented in this study to assist in interpretation of the flagged DIF items.

There is some disagreement about the effectiveness of the MH method. Meredith and Millsap (1992) found the MH to be inaccurate in detecting DIF, when the item responses were generated by complex IRT models, e.g., by a three-parameter logistic (3PL) model. On the other hand, Donoghue, Holland, and Thayer (1993) showed the MH method to have good Type-I error control and good power, even when the data were generated by complex IRT models.

Simultaneous Item Bias Test (SIBTEST). Another non-parametric test that has been developed for DIF detection is the SIBTEST (Shealy & Stout, 1993). Like the MH method, SIBTEST also provides a DIF statistic for detecting DIF as well as an estimate of the effect size of DIF. Originally, SIBTEST was only capable of detecting uniform DIF. Different from the MH method, however, is that SIBTEST can be used to detect whether DIF is present in one or more items simultaneously. To do this, two subtests need to be formed, one is the "suspect" subtest containing the item(s) suspected of functioning differentially and, therefore, to be tested for DIF, and the other is assumed to be the DIF-free subtest, containing items which are not suspected to function differentially. The score on the DIF-free subtest serves as the matching variable. A weighted mean difference in subtest performance between the focal group and the reference group, β_{UNI}, is computed as

$$\hat{\beta}_{UNI} = \sum_{k=0}^{K} p_k(\bar{Y}_{rk} - \bar{Y}_{fk}) \ , \tag{2.11}$$

where $\bar{Y}_{rk}$ and $\bar{Y}_{fk}$ are the mean scores on the suspect subtest for the reference group and the focal group, respectively, for a score k, $k = 0, \ldots, K$, on the DIF-free subtest. p_k is the proportion of examinees in the focal group with score k on the DIF-free subtest. The statistic for β_{UNI} is given by

$$B_{UNI} = \frac{\hat{\beta}_{UNI}}{\hat{\sigma}(\hat{\beta}_{UNI})} \ , \tag{2.12}$$

where $\hat{\sigma}(\hat{\beta}_{UNI})$ is the estimated standard error of $\hat{\beta}_{UNI}$. Shealy and Stout (1993) demonstrated the statistic B_{UNI} is approximately distributed as a standard normal (i.e., $N(0,1)$) under the null hypothesis, which is

$$H_0 : \beta_{UNI} = 0 \ . \tag{2.13}$$

Roussos and Stout (1996) provide standards for classifying DIF detected by SIBTEST based on the estimate of effect size $\hat{\beta}$:

A. Negligible DIF, where absolute value of $\hat{\beta} < .059$ and the hypothesis test is rejected.

B. Moderate DIF, where absolute value of $.059 \leq \hat{\beta} < .088$ and the hypothesis test is rejected.

C. Large DIF, absolute value of $\hat{\beta} \geq .088$ and the hypothesis test is rejected.

Although SIBTEST was initially described for detection of uniform DIF, it can also be used for detecting nonuniform DIF as well (Li & Stout, 1996).

2.2.2 Parametric DIF Methods

IRT is a family of statistical models relating the probability of a response on a test item to the latent ability measured by the test. The item characteristic curve (ICC) incorporates this information in the curve defined for a particular IRT model. IRT provides a natural framework within which to study DIF. In the IRT framework, DIF can be characterized as occurring when the ICC differs for the reference and focal groups. One approach to DIF detection can been seen as a task of comparing ICC's from the reference and focal groups.

Lord's Chi-Square. Since the ICC is completely determined by the item parameters, comparing ICC is same as comparing the item parameters from different groups. Lord (1980) developed a chi-square statistic for testing the equality of item parameters between the reference and focal groups. Lord's chi-square requires that the parameters of the IRT model be estimated separately for the reference and focal group. Next, the item parameters need to be placed on the same metric by means of some linking procedure. The null hypothesis for Lord's chi-square is

$$H_0 : \begin{cases} \mathrm{b}_f = b_r \\ \mathrm{a}_f = a_r \\ \mathrm{c}_f = c_r \end{cases} \tag{2.14}$$

To test this null hypothesis, Lord's chi-square is calculated by

$$\chi^2 = (a_{diff} b_{diff} c_{diff})' \Sigma^{-1} (a_{diff} b_{diff} c_{diff}), \tag{2.15}$$

where $a_{diff} = a_f - a_r, b_{diff} = b_f - b_r$, and $c_{diff} = c_f - c_r$, and Σ is the variance-covariance matrix of the differences between the parameter estimates. The statistic χ^2 is asymptotically distributed as a chi-square with p degrees of freedom, where p is the number of parameters being compared. Usually the guessing parameter c is not included for comparing since the standard error of c is typically large and will cause the test to be conservative.

Raju's Area Measures. An alternative way to compare ICCs is to calculate the area between the ICCs estimated in the reference and the focal group. In this method, the item parameters should be calibrated separately for the two groups and then a linking procedure should be used to placing the item parameters on a common scale. When that has been done, then the area between ICCs is calculated. An exact expression for computing the area between the ICCs for the common dichotomous IRT models was derived by Raju (1988). The expression under the 3PL model is

$$Area = (1-c)|\frac{2(a_f - a_r)}{Da_f a_r} ln[1 + e^{\frac{Da_f a_r (b_f - b_r)}{(a_f - a_r)}}] - (b_f - b_r)| \ . \tag{2.16}$$

For a two-parameter logistic (2PL) model, the term involving c disappears. For a one-parameter logistic (1PL) model, the expression reduces to the difference between the b values estimated from the two groups. Prior to Raju (1990), the disadvantage of this approach was that extra work needed to be done to establish an empirical cut-off value with which the area statistic could be compared to decide whether DIF was present. In this regard, Rogers and Hambleton (1989) suggested an approach using simulated data to establish a cut-off value. Raju (1990) provided a distribution for the signed and unsigned area measures, making it possible to do a statistical test for the significance of the group difference.

Likelihood Ratio Test for DIF. Another IRT-based method for DIF detection is the likelihood ratio test (Thissen, Steinberg, & Wainer, 1993). In this test, the likelihoods for two nested models are compared, a compact model and an augmented model. In the compact model, the item parameters for all items are assumed to be the same for both the reference and focal groups. In the augmented model, item parameters for all items except the studied item(s) are constrained to be equal in both the reference and focal groups. The likelihood ratio is calculated by

$$G^2 = -2\text{log}\left(\frac{L_C}{L_A}\right) \ , \tag{2.17}$$

where L_C and L_A are the likelihoods of the compact model and the augmented model, respectively. The statistic G_2 is distributed as a χ^2 under the null hypothesis with degrees of freedom p, where p is the difference in the number of parameters estimated in the compact and augmented models. The remaining items in the augmented model are constrained to have the same parameters for both the reference and focal groups. In addition, the remaining items serve as an anchor set to link the metrics of the focal and reference groups. Then DIF will be checked item by item.

2.2.3 Multidimensional DIF Detection

Most DIF detection procedures, whether parametric or non-parametric, are based on an assumption that the data are unidimensional. As a result, the matching variable conditions on a unidimensional raw score or latent ability score. Many cognitive ability tests are multidimensional to some extent, however, which requires a model that accounts for composites of several abilities (Ackerman, 1992). Ackerman (1992) and Ackerman and Evans (1994) suggested that a primary source of DIF is that due to the failure of the matching criterion to account for the complete latent ability space. In this regard, Walker and Beretvas (2001) examined DIF in a mathematics test from a multidimensional perspective. Results indicated some open-ended items measured mathematics ability as well as written mathematical communication, even though only a single score composed of multiple choice and open-ended items was used as the matching variable. Walker and Beretvas suggested that mathematic communication ability should be modeled in addition to mathematics ability.

If a test is essentially multidimensional, as Ackerman (1992) suggests, then matching groups according to a unidimensional criterion will result in more items flagged as DIF items, even though they may actually be measuring relevant aspects of the latent variable space. A more accurate measure of the latent ability might be helpful as the matching variable, in that it could help to reduce inflated estimates of DIF. In this regard, Ackerman and Evans (1993) demonstrated how DIF can be eliminated when two latent abilities were used in the matching variable.

Mazor, Hambleton, and Clauser (1998) compared results from MH and logistic regression methods for DIF detection using three different matching variables, total test score, relevant subtest score (i.e., the score from the subtest to which the studied item belongs), and all subtest scores. The matching criterion formed of all subtest scores was a multivariate composite of scores from all the subtests. A simulation study indicated that the multivariate subtest scores performed best in terms of reducing Type I errors. The relevant subtest score performed next best, followed by the total test score.

Zhang (2007) investigated DIF for CDMs. Since CDMs estimate mastery or non-mastery status for several attributes instead of for a single ability, Zhang used attribute mastery profiles estimated using the DINA model as the matching variables for both the MH and

SIBTEST methods. Use of attribute profiles instead of a single unidimensional score as the matching variable enabled Zhang to establish a measure of the ability space that had greater construct validity. Zhang's results indicated a reduced Type I error rate, compared with the use of a unidimensional score as the matching variable.

Clearly, the choice of an appropriate matching variable can improve the detection of DIF. One problem with the approach taken by Zhang (2007), however, is that the attribute mastery profiles for the reference and focal groups were estimated by the same DINA model. That is, the same item parameters were assumed for the reference and focal group. Once it is determined that DIF exists in some of the items on the test, the assumption is no longer tenable that the two groups share the same item parameters. The result of the violation of this assumption is that the estimates of attribute mastery profiles are biased, and the matching variable is contaminated. In addition, the MH and SIBTEST methods used by Zhang (2007) were only capable of detecting uniform DIF. Thus the power of DIF detection was low for items that were generated to have nonuniform DIF. In the present study, the model was calibrated with different sets of item parameters for the reference and focal groups. Next, group differences in both the guessing parameter and slip parameters were examined separately for each item to determine if DIF exists. In this way, it was possible to determine whether uniform or nonuniform DIF was present.

2.3 A Modified Higher-Order DINA Model for both DAF and DIF Analysis

In this study, a model-based method was developed for simultaneous detection of DAF and DIF simultaneously. We begin by making some adjustments to the higher-order DINA model:

- Templin (2004) proposed a generalized linear mixed proficiency space model (GLMPSM) that includes examinee covariates in the estimation of the proficiency space. In this study, we employ the GLMPSM, using a group indicator as an examinee covariate in the upper level of the higher-order DINA model. In the context of DIF detection, this group indicator was used as an index of either manifest group membership (e.g., ethnicity or gender) or cognitive differences (e.g., latent groups that differ with respect to use of cognitive strategies). A significance test of the coefficient estimated for the group indicator is described for detection of DAF.

- The number of attributes included on a test is usually small (e.g. 5). As a result, convergence and precision of attribute discrimination parameters may be problematic. As an example, Li et al. (2007) have shown that, when the number of items is small (e.g., 6), a mixture Rasch model (Rost, 1990) may converge, but more complex models such as the mixture 2PL and mixture 3PL may not. In the present study, therefore, the adjusted higher-order DINA model was estimated with the discrimination parameters fixed to be equal over all attributes and across groups. As a result, only uniform DAF was examined in this study.

- Item parameters, g and s, are located in the lower level of the higher-order DINA model. Values for these parameters were set to permit different values for different groups.

2.3.1 Attribute-Level Model Specification

Given the modifications noted above, the attribute level specification of the higher-order DINA model can be re-written as

$$\text{Logit}[P(\alpha_k|\theta_j)] = a(\theta_j + \Delta t I_j) - (\beta_k + \gamma_k I_j), \tag{2.18}$$

where

- I_j is a group indicator, that takes a value of 0, if examinee j belongs to the reference group, and 1, if examinee j belongs to the focal group;

- a is a uniform discrimination parameter, that is, it is fixed to be the same across attributes and groups;

- β_k is the difficulty parameter of attribute k for the reference group;

- γ_k is the difference in difficulty for attribute k between the reference and focal groups, and represents the amount of uniform DIF for attribute k. A positive sign for γ indicates the attribute favors the reference group, a negative sign indicates the attribute favors the focal group.
- θ_j is the general ability for member j of the reference group, and
- Δt is the mean difference in ability between the reference and focal group.

As noted above, in this model a is a common discrimination parameter for all attributes in both reference group and focal group. In addition, β_k is the attribute difficulty parameter for the reference group, $\beta_k + \gamma_k$ is the attribute difficulty parameter for the focal group, θ_j is the ability of examinee j, who belongs to reference group, and $\theta_j + \Delta t$ is the ability of examinee j, who belongs to focal group.

This model is a special case of a Rasch model with covariates, albeit with a discrimination parameter that does not necessarily equal 1. The model is not yet identified, however, since a constant (say, c) can be added to any β_k and the constant c/a can be added to any θ_j so that the odds ratio won't be changed. The same thing happens to γ_k and Δt. In order to solve the non-identifiability problem, the parameters in this study were adjusted using Chaimongkol's (2005) method for multilevel logistic regression models:

$$\beta_k^{adj} = \beta_k - \overline{\beta} \tag{2.19}$$

$$\gamma_k^{adj} = \gamma_k - \overline{\gamma} \tag{2.20}$$

$$\theta_j^{adj} = \theta_j - \overline{\beta}/a \tag{2.21}$$

$$\Delta t^{adj} = \Delta t - \overline{\gamma}/a \tag{2.22}$$

After the adjustment, DAF can be detected by examining if $\gamma_k^{adj} = 0$.

2.3.2 Item-Level Model Specification

And at the lower-level of the higher-order DINA model, the DINA model can be rewritten as

$$P(Y_{ij} = 1|\xi_{ij}) = (1 - s_{mi})^{\xi_{ij}} g_{mi}^{(1-\xi_{ij})} \ , \tag{2.23}$$

where s_{mi} is the slip parameter for item i in group m, and g_{mi} is the guessing parameter for item i in group m. Here m is r when examinee j belongs to the reference group, and f when examinee j belongs to focal group. Thus, the slip and guess parameters for item i in the reference group and the focal group are denoted as g_{ri}, s_{ri} and g_{fi}, s_{fi}, respectively.

As noted above, an item is said to be functioning differentially, when the probability of success on the item is higher for one group than for the other, even though examinees in both groups are matched on ability. In the DINA model framework, ξ_{ij} can be regarded as the ability variable with two levels: 1 indicates examinee j mastered all attributes required by item i, and 0 indicates examinee j missed at least one attribute required by item i.

By definition, in the DINA model, the conditional probability of a correct response to item i is $1 - s_i$, when $\xi = 1$, and g_i, when $\xi = 0$. Two straightforward ways to obtain an estimate of DIF are (1) to compare $1 - s_i$ and g_i for item i between the reference and focal groups, and (2) to marginalize the differences in the probability of success for item i across all levels of ability. The first way can be presented as

$$\Delta s_i = (1 - s_{fi}) - (1 - s_{ri}) = s_{ri} - s_{fi} \tag{2.24}$$

and

$$\Delta g_i = g_{fi} - g_{ri} \ , \tag{2.25}$$

where positive values of Δs_i indicate the item favors the reference group for examinees mastering all attributes required by item i, and positive values of Δg_i indicate the item favors the focal group for examinees who have not mastered at least one of the attributes required by item i. Thus, there are four combinations of Δs_i and Δg_i:

1. Both Δs_i and Δg_i are positive, that is, item favors the reference group for the masters but favors the focal group for the non-masters.

2. Both Δs_i and Δg_i are negative, that is, item favors the focal group for the masters but favors the reference group for the non-masters.

3. Δs_i is positive and Δg_i is negative, that is, item favors the reference group for both masters and non-masters.

4. Δs_i is negative and Δg_i is positive, that is, item favors the focal group for both masters and non-masters.

Combination 3 and 4 indicate uniform DIF, and combination 1 and 2 indicate non-uniform DIF.

The second way can be presented as

$$D_i = [(1 - s_{fi}) - (1 - s_{ri})] \times P(\xi_i = 1) + (g_{fi} - g_{ri}) \times P(\xi_i = 0) \ . \tag{2.26}$$

An important problem with this second approach, however, is that cancellation of the DIF effect could occur, when the group with the higher $1 - s_i$ has a lower g_i, or when the group with a lower $1 - s_i$ has a higher g_i. DIF cancellation is an undesirable outcome as it can potentially mask adverse DIF. As a result, we obtained a DIF index using the first approach. That is, DIF was detected in this study by testing whether $\Delta g_i = 0$ or $\Delta s_i = 0$.

Chapter 3

Research Designs and Methods

3.1 A Markov chain Monte Carlo Algorithm for Model Estimation

A Markov chain Monte Carlo (MCMC) algorithm employing Gibbs sampling was used to estimate the model parameters in this study. This algorithm is implemented in the WinBUGS software (Spigelhalter, Thomas, & Best, 2003) and was used to simulate a Markov chain in which values representing parameters of the model are repeatedly sampled from their full conditional posterior distributions over a large number of iterations. In this way, the MCMC algorithm can be used to sample values in each iteration for each of the parameters in the model conditional on those parameters already estimated up to that point in the iteration. With respect to the higher-order DINA model, model parameters were sampled from their full posterior distributions conditional upon the already sampled ability and examinee attribute mastery parameters.

To derive the posterior distributions for each parameter, it is first necessary to specify their prior distributions. The following priors were used to estimate the parameters of the modified higher-order DINA model in this study: $\alpha_{jk} \mid \theta_j \sim \text{Bernoulli}(a(\theta_j + \Delta t I_j) - (\beta_k + \gamma_k I_j))$, $\theta_j \sim N(0,1)$, $a \sim N(0,1)$ $a > 0$, $\beta_k \sim N(0,100)$, $\gamma_k \sim N(0,100)$, $\Delta t \sim N(0,1)$, $g_{mi} \sim \text{Beta}(U_g,\, S_g)$, $s_{mi} \sim \text{Beta}(U_s,\, S_s)$, $U_g \sim \text{Uniform}(.1,.9)$, $S_g \sim \text{Uniform}(.5,10)$, $U_s \sim \text{Uniform}(.1,.9)$, $S_s \sim \text{Uniform}(.5,10)$.

The beta distribution used as a prior for the g_{mi} and s_{mi} parameters was intended to ensure that the ranges of g_{mi} and s_{mi} were between 0 and 1, respectively. When the value of U_g (or U_s) was larger than the value of S_g (or S_s), the priors for g and s were more likely to be drawn from the range of .5 to 1, otherwise, the priors for g_{mi} and s_{mi} would be more likely to be drawn from the range of 0 to .5. The uniform distribution (.1, .9) for U_g (or U_s) and uniform distribution (.5, 10) for S_g (or S_s) as the hyperprior distribution helped ensure that S_g (or S_s) was more likely to be larger than U_g (or U_s) with the result that g_{mi} and s_{mi} were more likely to be drawn from between 0 and .5. In this way, the beta distribution is a more informative and realistic prior distribution than a uniform distribution (0, 1).

Some traps occurred in running the MCMC algorithms after applying the above priors. Step-by-step checking found these traps could be stopped by reducing the variances on the priors for β and γ from 100 to 1. The resulting $N(0, 1)$ priors were finally used for estimating β and γ.

One benefit of the MCMC algorithm implemented in this model-based DIF analysis is that the MCMC algorithm estimates a posterior distribution for all sampled parameters, including the DAF and DIF indices. Further, the posterior distribution for the parameter provides a $100(1-\alpha)\%$ credibility interval that can be used to examine if the magnitude of DIF equals 0 (Samuelsen, 2005). In this test, DIF is present, if the interval does not contain 0. Thus, we can judge whether DAF exists for attribute k by testing, if the $100(1-\alpha)\%$ credibility interval on γ_k^{adj} contains 0. Similarly, whether DIF exists for item i can be determined by testing if the $100(1-\alpha)\%$ credibility intervals on Δg_i or Δs_i contain 0.

3.2 A Simulation Study

This simulation study has two purposes. First, it is designed to determine whether the new model-based method can be used to detect both DAF and DIF simultaneously. Second, it is designed to compare the Type I error control and power of this approach to the usual methods for DIF detection or to the non-model based methods for DIF detection in cognitive diagnostic assessment.

We begin by first constructing conditions. For this purpose, the first task of this simulation study is to build DAF and DIF conditions for the higher-order DINA model. We do this by manipulating several factors and use these to generate the response data for the models.

3.2.1 Data Simulation Design

Factors Potentially Affecting Detection of DIF and DAF. The performance of traditional DIF detection methods has been shown to be affected by factors such as sample size, test length, proportion of items on the test containing DIF, the amount of DIF, ability distribution difference between the reference and focal group (Mazor, Clauser, & Hambleton, 1992; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). In addition to these factors, other factors also may be appropriate to consider when the context is one which considers both DIF and DAF detection. These additional factors could include the number of attributes included in the model for a fixed test length, the completeness and correctness of the Q-matrix specification, the complexity of Q-matrix, the correlations among attributes, and the values for the slip and guessing parameters. The completeness of a Q-matrix refers to whether or not all attributes required to answer the items correctly are specified. Completeness also requires that each attribute is measured by at least one item. The correctness of a Q-matrix means that the specification of the elements of the matrix correctly indicate which attributes are required for correctly answering each item. Finally, the complexity of a Q-matrix refers to the ratio of the number of items to the number of attributes. In this study, we manipulated the complexity of a Q-matrix.

It is not feasible to examine the impact of all the factors noted above in this first study of this new model. In this study, therefore, only a portion of the factors was manipulated and the remaining ones fixed.

A single, 25-item test with five attributes was simulated. In addition, a 1PL IRT model was used as the IRT model in the higher-order DINA model to estimate how general ability and attribute difficulties predict mastery status of attributes. This is analogous to examining how ability and item difficulty predict the correct response to an item in the usual IRT framework. It is well-known that the precision of ability estimates increases with an increase in number of items. Similarly, more attributes should lead to improvement of the precision of model attribute parameters. However, in practice, the number of attributes to be measured in a test is typically small. In this study, therefore, five attributes were simulated for the 25 items.

Previous research has suggested the percentage of DIF items has an impact on DIF detection. Too many DIF items can contaminate the conditioning variables (Gierl, Gotzmann, & Boughton, 2004; Narayanon & Swaminathan, 1996). Narayanon and Swaminathan (1996) simulated data with DIF in up to 20% of the items. Likewise, Zhang (2007) simulated five DIF items on a 25-item test. The percentages of differentially functioning items and attributes to be studied is determined based on rates that might occur in actual testing programs. One problem with this approach is that most well-developed tests have been carefully constructed so that most DIF has been removed from the test. Even so, Mazor, Kanjee, and Clauser (1995) found approximately 20% items of the items functioned differentially on an operational test. With respect to attribute functioning, no research has yet been reported about the prevalence of DAF. Consequently, in this study we selected five of the 25 items (i.e., 20%) as the single DIF condition. Further, two of the five attributes were manipulated to provide DAF conditions so that it was possible to simulate one attribute discriminating the focal group and another one favoring the focal group.

In the simulation study and the real data example presented in de la Torre and Douglas (2004), the range of attribute difficulty was from -1.5 to .5, and the range of most slip and guessing parameters was from .1 to .3. These ranges were incorporated in this study such that five attribute difficulty parameters for the reference group were fixed at (0, 0, 0, 0, 0). Slip and guessing parameters for the 25 items were generated from a uniform distribution between .1 and .3. This meant that a master had a probability of from .1 to .3 of responding incorrectly to the item, whereas a non-master had a probability of from .1 to .3 of responding correctly. These values of guessing and slip parameters are reasonable for a test with good diagnostic quality. Since slip and guessing parameters are both indicators of noisiness in the data, smaller values for these parameters would indicate that the model is likely to be more diagnostically useful for distinguishing masters and non-masters (Templin & Henson, 2006).

For the two differentially functioning attributes and the five differentially functioning items simulated in this study, the amount of DIF and DAF was also fixed. Zhang (2007) set two levels of DIF for item slip and guessing parameters: .075 and .15. A DIF of .075 was not sufficiently large enough to be detected, although a DIF of .15 yielded a power of 1.0 for some conditions in Zhang's study (2007). The amount of DIF, however, was not the focus of this study. Consequently, the level of DIF selected for this study was .1, a value between .075 and .15. That is, the slip and guessing parameters in the focal group were formed by adding or subtracting .1 from the values for the reference group.

No research has yet been reported describing the amount of DAF that is reasonable to expect in a practical testing situation. In part, this is because CDMs are only a relatively recent addition to the psychometric literature, and DAF has yet to be studied in this context. In this study, therefore, DAF was simulated as 1.0 between attributes. This is a relatively large difference. The attribute difficulty in the focal group, in other words, was simulated as $(1, -1, 0, 0, 0)$. This pattern simulates the first attribute as harder and the second attribute as easier than the same two attributes in the reference group. The remaining three of the five attributes were equally difficult for members of the reference and focal groups.

Factors to be Manipulated. The factors described above were fixed in this study. The following five factors were manipulated: 2 sample sizes, 2 ability distributions, 2 Q-matrices with different levels of complexity, 3 different patterns of attribute discrimination parameters, and 6 scenarios with different combinations of DIF and DAF.

Sample size is consistently shown to be an important factor in DIF detection studies. Previous research has shown that DIF detection using the MH and the LR statistics improve as sample size increased (Mazor, Clauser, & Hambleton, 1992; Narayanon & Swaminathan, 1996; Swaminathan & Rogers, 1990). Since a model-based method is proposed in this study, sample size is likely to be important. de la Torre and Douglas (2004) report good recovery for the one-group higher-order DINA model with a sample size of 1,000. In this study, therefore, 1,000 examinees for the reference group and 1,000 examinees for the focal group were simulated as one sample size condition. A second sample size of 500 examinees per group was simulated as a small sample condition.

Group differences in ability have been shown to affect DIF detection (Mazor, Clauser, & Hambleton, 1992; Shealy & Stout, 1993; Narayanon & Swaminathan, 1996). Unmatched ability distributions make DIF detection more difficult than matched ability distributions. For this reason, two ability distributions were simulated in this study: A matched ability distribution situation was simulated in which both the reference and focal group had the same ability distribution of $N(0, 1)$, and an unmatched ability condition was simulated, in which the reference group ability distribution was simulated as $N(0, 1)$, and the focal group was simulated as $N(-1, 1)$.

In CDMs such as the higher-order DINA model, the Q-matrix is used to link items and attributes. Because of its centrality in the model, a misspecified Q-matrix can result in poor estimation of model parameters. One problem with the specification of the Q-matrix is that as yet, no research has been reported in which the Q-matrix has been used to guide item and test development. As a result, psychometric research with CDMs has relied on content expert judgments to determine the attributes measured by existing items on existing tests that were developed using other methods. Consequently, it is likely that the Q-matrices thus specified are not as effective as they might be were the items and tests developed to specifically measure these attributes.

Further, a trade-off has been shown to exist between the complexity of the Q-matrix and parameter estimation accuracy for a fixed test length (Hartz, 2002; Zhang, 2007). When the items measure too few attributes, the information to estimate the model parameters is not sufficient and estimation suffers. In contrast, when each item measures too many attributes, the capability to distinguish between attributes by the model is questionable. In this study, a single 25×5 Q-matrix was constructed that was balanced between complexity and effectiveness. For first five items, each item was simulated as estimating a single attribute; for Item 6 to 15, each item was simulated as estimating two attributes; for Item 16 to 25,

Table 3.1: Q-Matrix with Complex Structure

	Attributes						Attributes						Attributes				
Items	1	2	3	4	5	Items	1	2	3	4	5						
1	1	0	0	0	0	10	0	1	1	0	0	18	1	1	0	0	1
2	0	1	0	0	0	11	0	1	0	1	0	19	1	0	1	1	0
3	0	0	1	0	0	12	0	1	0	0	1	20	1	0	0	1	1
4	0	0	0	1	0	13	0	0	1	1	0	21	1	0	1	0	1
5	0	0	0	0	1	14	0	0	1	0	1	22	0	1	1	1	0
6	1	1	0	0	0	15	0	0	0	1	1	23	0	1	1	0	1
7	1	0	1	0	0	16	1	1	1	0	0	24	0	1	0	1	1
8	1	0	0	1	0	17	1	1	0	1	0	25	0	0	1	1	1
9	1	0	0	0	1												

each item was simulated as estimating three attributes. This Q-matrix design is shown in Table 3.1. As can be seen in Table 3.2, each attribute is measured by 11 items. This design is similar to that of de la Torre and Douglas (2004) for a 30-item test. The Q-matrix specified by de la Torre and Douglas, however, repeated the one-attribute-by-one-item pattern for 5 items twice.

In addition, a second Q-matrix was specified for this simulation study (see Table 3.2). In this Q-matrix, a 25 × 5 Q-matrix was specified that has what is known as simple structure. In a simple structure Q-matrix, each item is modelled as measuring only a single attribute. The Q-matrix in Table 3.1 shows that each attribute was measured by five items. This is not necessarily the most desirable form of Q-matrix, but it is common in existing tests. As an example, most statewide tests only measure a single content strand per item. This is essentially the same as measuring a single attribute per item. The purpose of using two Q-matrices with different complexity was to compare whether the number of items per attribute had a differential impact on parameter estimation and, further, whether the accuracy of DIF or DAF were affected.

Table 3.2: Q-Matrix with Simple Structure

	Attributes						Attributes						Attributes				
Items	1	2	3	4	5	Items	1	2	3	4	5	Items	1	2	3	4	5
1	1	0	0	0	0	10	0	1	0	0	0	18	0	0	0	1	0
2	1	0	0	0	0	11	0	0	1	0	0	19	0	0	0	1	0
3	1	0	0	0	0	12	0	0	1	0	0	20	0	0	0	1	0
4	1	0	0	0	0	13	0	0	1	0	0	21	0	0	0	0	1
5	1	0	0	0	0	14	0	0	1	0	0	22	0	0	0	0	1
6	0	1	0	0	0	15	0	0	1	0	0	23	0	0	0	0	1
7	0	1	0	0	0	16	0	0	0	1	0	24	0	0	0	0	1
8	0	1	0	0	0	17	0	0	0	1	0	25	0	0	0	0	1
9	0	1	0	0	0												

Shealy and Stout (1993) and Roussos and Stout (1996) have noted that the primary cause of DIF is due to unmodelled multidimensionality, specifically to that dimensionality which is not relevant to the construct being measured. In this study, we distinguished between construct-relevant multidimensionality and construct-irrelevant multidimensionality. DIF was simulated as construct-irrelevant multidimensionality. Items with construct-relevant multidimensionality were not considered as DIF items. Unfortunately, the standard DIF analysis approaches in which conditioning is done based on a unidimensional model of ability can not be used for detection of this type of DIF.

In the context of cognitive diagnostic models, however, this is not a problem. When attributes are highly correlated, the test is still unidimensional and items will not be detected as DIF using standard methods. Whereas, when correlations among attributes are low, the test is no longer unidimensional and items will be detected as DIF by standard methods. In order to compare the standard DIF analyses and the method proposed in this paper under the different levels of attribute correlation, the attributes were generated with three correlation levels. These levels were .2, .4, and .8, and were considered to be low, moderate, and high levels of correlation.

In the higher-order DINA model, the attributes are estimated by the IRT portion of the model. The correlations among attributes can be manipulated by the discrimination parameters in the model. As mentioned previously, non-uniform DAF was not simulated in this study. In terms of the simulation conditions, this means that the attribute discrimination parameters was constrained to be equal over all attributes across groups, so only a single value was assigned to generate discrimination parameters. Results from a small simulation study indicated that, when the discrimination parameter equaled 1, 2, or 6, the 1PL higher level portion of the model could accurately estimate attributes, when correlations were .2, .4 and .8 respectively. Discrimination parameters as high as 6 do not seem practical for most items. However, attributes in an operational test are sometimes highly correlated. That is, the person mastering one attribute is more likely to master other attributes. For example, the correlation among the content strands was around .7 in the Florida Comprehensive Achievement Test (FDOE, 2003) statewide mathematics test.

In summary, the following conditions were considered with respect to combinations of DAF and DIF.

1. No DAF and DIF. In this condition, the attribute parameters, the item slip and guessing parameters were the same for the focal group and for the reference group.

2. DAF only. Attribute difficulty parameters for the focal group were set as $(-1, 1, 0, 0, 0)$. This means that the first attribute was easier for the focal group, and the second attribute was harder for the focal group. The three remaining attributes were simulated to be equally difficult for both the reference and focal groups.

3. Uniform DIF only. The five DIF items were generated by increasing the guessing parameters by .1 and decreasing the slip parameters by .1 for the five DIF items for the focal group relative to the reference group. The non-DIF items were simulated with the same

item parameters as the reference group. For the slip and guessing parameters, the probability of an item for a master equals the guessing parameter g_i and the probability of success on an item for a non-master equals $1 - s_i$. Under this condition, both masters and non-masters in the focal group were simulated to have probabilities of .1 higher than those in the reference group.

4. Non-uniform DIF only. The slip and guessing parameters were simulated in the five DIF items by increasing these parameters by .1 in the focal group. That means masters in the focal group were simulated as having lower probabilities for answering the questions correctly than those in the reference group, whereas non-masters in the focal group were simulated as having higher probabilities of getting the same item correct than those in the reference group. In this study, Non-uniform DIF is similar to DIF in IRT in which item characteristic curves for the reference and focal groups crossed. It was not similar, however, to the situation in which DIF occurs only due to different item discriminations but the item characteristic curves do not actually cross.

5. DAF and uniform DIF. Attribute difficulty parameters for the focal group were set as $(-1, 1, 0, 0, 0)$ so that the focal group had a positive DAF for first attribute and a negative DAF for second attribute. Simultaneously, at the lower level, the slip and guessing parameters of the five DIF items were simulated by decreasing the slip parameter by .1 and increasing the guessing parameter by .1 for the focal group.

6. DAF and non-uniform DIF. Attribute difficulty parameters for the focal group were also set as $(-1, 1, 0, 0, 0)$. Both the slip and guessing parameters of the five DIF items were simulated by increasing by .1 for the focal group.

In all, five factors were manipulated in this simulation study: 2 levels of sample size, 2 levels of ability distribution difference, 2 kinds of Q-matrix with different complexity, 3 levels of attribute discrimination parameters (i.e., by the three levels of attribute correlations), and

6 scenarios of DIF and DAF (shown above). Therefore, $2 \times 2 \times 2 \times 3 \times 5 = 120$ conditions were simulated, and 25 replications were done for each simulated condition.

3.2.2 Data Simulation Procedures

The following general data simulation procedures were used:

1. Construct two sets of attribute difficulty parameters, $(0, 0, 0, 0, 0)$ for the reference group and $(0, 0, 0, 0, 0)$ for the focal group in No DAF scenarios, and $(-1, 1, 0, 0, 0)$ for the focal group in the scenarios with DAF (See Table 3.3). Fix the same discrimination parameter (1, 2 or 6) for all attributes in both groups.

2. Randomly generate ability parameters for each replication as $N(0, 1)$ for the reference group and $N(-1, 1)$ for the focal group.

3. Simulate the mastery profiles for all examinees in both groups based on the 1PL model, using examinee abilities as generated from 2 (above) and different sets of attribute parameters from 1 (above).

4. Randomly generate the slip and guessing parameters for the reference group from uniform (.1, .3) only once. Keep guessing and slip parameters for the reference group the same across different scenarios, and manipulate them for the focal group based on different scenarios according to the description of the six combinations of DAF and DIF given above (see Table 3.4). Keep all slip and guessing parameters for the two groups the same across the replications in the same condition.

5. Generate examinee responses based on the DINA model, by using examinee mastery profiles as in 3 (above) and using the two sets of slip and guessing parameters shown in 4 (above).

Table 3.3: Generating Attribute Parameters in Different Scenarios

	β's	γ's
No DAF	(0,0,0,0,0)	(0,0,0,0,0)
With DAF	(0,0,0,0,0)	(-1,1,0,0,0)

Table 3.4: Generating Item Parameters in Different Scenarios

	Reference Group		Focal Group					
			No DIF		Uniform DIF		Non-uniform DIF	
Item	g	s	g	s	g	s	g	s
1	**.15**	**.26**	**.15**	**.26**	**.25**	**.16**	**.25**	**.36**
2	.25	.16	.25	.16	.25	.16	.25	.16
3	.29	.16	.29	.16	.29	.16	.29	.16
4	.19	.21	.19	.21	.19	.21	.19	.21
5	.16	.23	.16	.23	.16	.23	.16	.23
6	**.24**	**.15**	**.24**	**.15**	**.34**	**.15**	**.34**	**.25**
7	.26	.23	.26	.23	.26	.23	.26	.23
8	.26	.29	.26	.29	.26	.29	.26	.29
9	.28	.18	.28	.18	.28	.18	.28	.18
10	.26	.23	.26	.23	.26	.23	.26	.23
11	**.28**	**.24**	**.28**	**.24**	**.38**	**.14**	**.38**	**.34**
12	.18	.21	.18	.21	.18	.21	.18	.21
13	.25	.24	.25	.24	.25	.24	.25	.24
14	.11	.23	.11	.23	.11	.23	.11	.23
15	.19	.29	.19	.29	.19	.29	.19	.29
16	**.14**	**.15**	**.14**	**.15**	**.24**	**.15**	**.24**	**.25**
17	.23	.25	.23	.25	.23	.25	.23	.25
18	.21	.21	.21	.21	.21	.21	.21	.21
19	.23	.21	.23	.21	.23	.21	.23	.21
20	.28	.23	.28	.23	.28	.23	.28	.23
21	**.22**	**.24**	**.22**	**.24**	**.32**	**.14**	**.32**	**.34**
22	.24	.27	.24	.27	.24	.27	.24	.27
23	.24	.12	.24	.12	.24	.12	.24	.12
24	.29	.18	.29	.18	.29	.18	.29	.18
25	.28	.29	.28	.29	.28	.29	.28	.29

The five DIF items are bolded

3.2.3 Convergence

Convergence diagnostics are used to help determine the number of iterations that should be used for burn-in and the number of post-burn-in iterations that should be used to estimate the posterior distribution. The Gelman and Rubin (1992) convergence diagnostic was used as implemented in the software BOA (Smith, 2005). The index of this diagnostic, $\hat{R}$, is estimated based on a comparison of the within and between chain variance for each variable. If $\hat{R}$ is approximately equal to 1 (or, as a rule of thumb, the 0.975 quantile is less than 1.2), the sample is considered to have reached a stationary distribution. Each condition in the simulation study was run with three parallel chains with over-dispersed initial values. Results indicated convergence was obtained after 1,000 iterations (i.e., the criterion $\hat{R} < 1.2$ was satisfied after 1000 iterations) for all structure parameters. Thus, a conservative burn-in of 4,000 iterations and 10,000 post burn-in iterations were used in all conditions. The examples of trace plots as well as the plots for the Gelman and Rubin statistics are given in Appendix B for two selected conditions: (1) Complex structure, 500 examinees per group, $a = 6$, unmatched ability distribution, both DAF and non-uniform DIF, and (2) Simple structure, 1000 examinees per group, $a = 1$, matched ability distribution, both DAF and uniform DIF.

The MCMC chain for each replication, in other words, was run for a total of 14,000 iterations. The amount of time required for each MCMC chain to run to completion differed depending on the data being analyzed. The MCMC run for a data set with 500 examinees per group required about 10 hours for completion on an HP BL460c 2.00 GHz server blade with a Quad-Core Intel Xeon processor and 5GB RAM running a Windows 2003 server operating system. The time required for running 14,000 MCMC iterations for a data set with 1,000 examinees per group on this same computer system was about 19 hours.

3.2.4 Recovery Analysis for the Modified Higher-order DINA Model

Before proceeding with the DIF and DAF detection portion of the simulation study, a recovery analysis was conducted to determine the extent to which the generating parameters could be recovered from the simulated data sets by the modified higher-order DINA model. The recovery analysis considered three issues, recovery of the simulated item parameters (i.e., the slip and guessing parameters), recovery of the simulated attribute difficulty parameters, and recovery of the attribute mastery classifications. Recovery of item parameters or attributes difficulty parameters was assessed using root mean squared errors (RMSEs) between the generating parameters and the parameter estimates. The RMSEs can be expressed as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{b_i} - b_i)^2}\ , \tag{3.1}$$

where b_i is the generating parameter for either an item or an attribute parameters, $\hat{b_i}$ is the parameter estimate, and n is the number of items or attributes. Recovery of attribute mastery classification was done by simply calculating the proportion of examinees who are correctly classified as masters or non-masters on each attribute.

Since the item parameters (i.e., the slip and guessing parameters) are both on the probability scale, which is invariant, no extra work needs to be done to transform the estimated parameter to the same scale with the generating parameters. The estimates for attribute difficulty parameters were first transformed to a common metric using the mean and sigma method (Loyd & Hoover, 1980) before RMSEs were calculated. Because attribute difficulty was adjusted to have a mean of zero for model identification, and the mean of generated attribute difficulty also equaled zero in the simulation design, these parameters were on the same metric and did not need transformation.

3.2.5 DAF and DIF Analysis Procedures

After the data were simulated, the following analyses were done to detect both DAF and DIF using the model-based method proposed in this dissertation. At the same time, the DIF analysis of the model based method was compared with the MH method with total scores as a matching criterion and with the MH method with attribute mastery profiles as a matching criterion as developed by Zhang (2007).

1. Run the modified higher-order DINA model for simulated response data in WinBUGS to obtain adjusted γ, Δg, Δs and their corresponding $100(1-\alpha)\%$ CI, simultaneously save the attribute mastery profiles for all examinees.
2. Detect DAF by checking against the $100(1-\alpha)\%$ credibility interval (CI) to see if the interval on γ contains 0.
3. Detect DIF by
   - checking if the $100(1-\alpha)\%$ CI for either Δg or Δs contains 0,
   - computing MH statistics using the total score as the matching criterion.
   - computing MH using attribute mastery profile as matching criterion,
4. Calculate the Type I error and power for all above DAF and DIF analysis procedures over all replications under each condition.

Since power is only assessed when the Type I error is controlled, Type 1 error control needs to be evaluated first. Type I error control was examined at the $\alpha = .05$ level. Bradley's (1978) liberal criterion of a range from .025 to .075 for Type I error rate for a nominal rate of .05 was used as the criterion in this study. To assess empirical Type I error control for DAF, the three non-DAF attributes in the scenarios with DAF and all five attributes in the scenarios without DAF were investigated by examining the percentage of these non-DAF attributes mistakenly detected as DAF. In the same way, to assess the empirical Type I error control for DIF, 20 non-DIF items in the scenarios with DIF and all 25 items in the scenarios without DIF were investigated by examining the percentage of these non-DIF items mistakenly detected as DIF. Only the Type I error rate for DAF was assessed for the new method developed in this study, although the Type I error rate for DIF was assessed and compared for both the new method and the MH methods based on the two different matching criteria. Power was assessed for those conditions in which the Type I error rate was controlled at the nominal level of significance. To assess the empirical power for DAF, the two DAF attributes in the scenarios with DAF were investigated by examining the percentages of these DAF attributes correctly detected as DAF. To assess the empirical power for the DIF, the five DIF items in the scenarios with DIF were investigated by examining the percentage of these items correctly detected as DIF. Similarly, only the power for DAF was assessed for the new method.

All Type I error and power analyses were calculated for each replication, then averaged over all replications under each condition. After these results were obtained, both Type I error rates and power were compared among conditions and among the DIF methods. Therefore, the following questions were addressed based on the results:

1. If the model-based method maintains Type I error control, what is its power for simultaneous detection of both DAF and DIF? How do sample size, ability distribution differences, Q-matrix specification, and attribute discrimination affect the performance of DAF and DIF detection? Is DIF detection for the data with non-uniform DIF as powerful as that for the data with uniform DIF?

2. Does MH based on total score maintain Type I error control? Is this control a function of the attribute discrimination parameter and weather DAF is present? If errors occur, do they occur only for those items measuring the DAF attributes?

3. How is the performance of DIF analysis by the model based method different from the MH using attribute profiles as the matching criterion?

4. For non-uniform DIF scenarios, does the new method perform better than both MH-methods?

Chapter 4

Results

The results of this study are presented in three parts: In the first two parts, results for the simulation study are presented; in the third part, a real data application is discussed. The simulation study includes the results for differential attribute functioning (DAF) detection and the results for differential item functioning (DIF) detection. Since DAF and DIF detection in this study are both model-based, the performance of DAF and DIF detection largely relies on how well the relevant parameters are recovered. Therefore, in the following discussion, the recovery of DAF relevant parameters or DIF relevant parameters is presented first. Then the Type I error and power of DAF are reported in the DAF part, and Type I error and power of DIF, for both DIF-g and DIF-s, are reported in the DIF part. Since the Mantel-Haenszel (MH) method is only designed to detect DIF, the comparison between the method presented in this study and the MH methods based on total scores and on attribute profiles is limited to the discussion of DIF. That is, Type I error control and power of DIF were compared between the model-based method presented in this study and the MH method using the two different matching criteria. As described in Chapter 3, recovery, Type I error control and power were evaluated under varied testing conditions with the combination of the following factors: Q-matrix structure, attribute discrimination parameters, sample size, ability distribution difference, scenarios of DIF and DAF combination.

4.1 DAF Detection

4.1.1 The Recovery of Higher-Level Parameters

Higher-level parameters mainly include attribute discrimination, a, attribute difficulty of the reference group, β, and group difference in attribute difficulty, γ. The recovery of the these three parameters was evaluated by RMSE. The recovery results for these three parameters under all testing conditions is presented in Tables 4.1 to 4.3. Table 4.4 summarizes the recovery information in Tables 4.1 to 4.3 for each parameter under each level of the five simulated factors.

In Table 4.1, it can be seen that recovery was generally good for the discrimination parameter, when $a = 1$, and almost as good when $a = 2$, under most conditions for either simple or complex structure Q-matrices. Recovery was less accurate, however, when $a = 6$ for both simple and complex structure. This latter result appeared to be more pronounced for the unmatched ability condition with the small sample size (i.e., N = 500 examinees per group) under both simple and complex structure. This can be seen in the relatively high RMSEs for a parameter, which are 1 or above. (The original estimates of a, prior to estimating RMSEs, revealed shrinkage to 5 or lower when the generating value was 6.) This is consistent with the result of the simulation study in de la Torre and Douglas (2004), in which the higher level with 8 attributes and one discrimination parameter shrank to around 4 from the generating value of 4.97. The small number of attributes in the present study could be one reason for the somewhat larger bias in the a parameter. In this regard, as in IRT models, the recovery of attribute discrimination may also be less accurate with only five attributes.

In Table 4.2, the RMSEs for attribute difficulty ranged from .07 to .20. This indicates generally good recovery. Recovery of β did not appear to be affected much by any of the different scenarios (i.e., combinations of the conditions in the simulation design) or whether the ability distributions were matched or not matched between the reference and focal groups.

This may have occurred because the attribute difficulty parameters for the reference group were all generated as 0 for all simulation conditions. Only the attribute difficulty, item guessing and slip parameters for the focal group were simulated to be different. Moreover, for both matched and unmatched ability distributions, that for the reference group was always generated as $N(0, 1)$.

Results in Table 4.3 indicated that the RMSEs of γ (i.e., the DAF parameter which indicated the group difference in attribute difficulty) were clearly higher than those for β. This was consistent with results from Chaimongkol (2005) in which a two-level DIF detection model was compared with the higher level DAF detection model of this study. In that study, the difference was that discrimination was constrained to be 1 in model estimation. In addition, RMSEs for recovery of β were around .10, and RMSEs for γ around .17, for $a = 1$, a sample size of 1,000 and matched ability distributions. This was comparable with the results in Tables 4.2 and 4.3 for the conditions $a = 1$, sample size = 500 examinee per group, matched ability distribution under simple structure.

As can be seen in Table 4.4, a similar pattern was found in the recovery of each of the three attribute parameters: The conditions with simple structure Q-matrix, lower attribute discrimination, and larger sample size produced smaller RMSEs for a, β, and γ. Results for different DAF and DIF combinations, however, did not show clear differences in RMSEs for a, β, and γ. The only difference was that recovery of a and γ appeared to be better in the matched ability condition, but not for the recovery of β. (See explanation in the above paragraph for the description about Table 4.2.) Of all three parameters, the recovery of γ, the group difference in attribute difficulty, is most relevant to the performance of DAF. Recall that DAF was determined by checking against the 100(1- α)% credibility interval to see whether the interval of γ contained 0 (here α refers to the significance level).

Table 4.1: RMSE of Attribute Discrimination a over 25 Replications

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	No DIF and DAF	.08	.11	.64	.05	.23	.81
		DAF only	.10	.17	.62	.06	.11	.86
		Uniform DIF only	.10	.15	.64	.06	.12	.77
		Non-uniform DIF only	.13	.16	.67	.06	.11	.84
		DAF and uniform DIF	.08	.09	.70	.06	.12	.85
		DAF and non-uniform DIF	.11	.11	.66	.06	.12	.89
	unmatched	No DIF and DAF	.10	.16	1.19	.08	.24	1.20
		DAF only	.13	.16	1.10	.09	.11	1.24
		Uniform DIF only	.08	.17	1.08	.09	.13	1.29
		Non-uniform DIF only	.09	.17	1.08	.08	.11	1.32
		DAF and uniform DIF	.09	.17	1.10	.09	.12	1.22
		DAF and non-uniform DIF	.10	.17	1.11	.09	.12	1.26
1000/g	matched	No DIF and DAF	.09	.05	.65	.09	.14	1.08
		DAF only	.08	.09	.57	.06	.07	1.06
		Uniform DIF only	.10	.04	.66	.09	.14	1.08
		Non-uniform DIF only	.10	.05	.67	.05	.13	1.11
		DAF and uniform DIF	.07	.05	.59	.06	.13	1.01
		DAF and non-uniform DIF	.04	.05	.60	.06	.13	1.05
	unmatched	No DIF and DAF	.04	.08	.67	.05	.06	1.00
		DAF only	.09	.08	.72	.08	.09	1.04
		Uniform DIF only	.10	.08	.67	.05	.07	1.00
		Non-uniform DIF only	.10	.08	.66	.08	.07	1.02
		DAF and uniform DIF	.04	.08	.72	.07	.10	1.04
		DAF and non-uniform DIF	.04	.08	.75	.08	.10	1.04

Table 4.2: RMSE of Attribute Difficulty β over 25 Replications

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	No DIF and DAF	.12	.13	.20	.16	.14	.18
		DAF only	.12	.13	.20	.16	.15	.18
		Uniform DIF only	.12	.16	.20	.16	.14	.18
		Non-uniform DIF only	.12	.16	.20	.16	.14	.18
		DAF and uniform DIF	.12	.13	.19	.16	.15	.12
		DAF and non-uniform DIF	.12	.13	.19	.16	.15	.17
	unmatched	No DIF and DAF	.12	.12	.19	.17	.14	.17
		DAF only	.12	.12	.19	.16	.15	.18
		Uniform DIF only	.12	.12	.19	.17	.15	.17
		Non-uniform DIF only	.12	.12	.19	.17	.15	.17
		DAF and uniform DIF	.12	.12	.19	.16	.15	.17
		DAF and non-uniform DIF	.12	.12	.19	.16	.16	.17
1000/g	matched	No DIF and DAF	.08	.09	.14	.10	.12	.18
		DAF only	.07	.09	.14	.13	.13	.14
		Uniform DIF only	.07	.09	.14	.10	.13	.18
		Non-uniform DIF only	.07	.09	.14	.10	.13	.18
		DAF and uniform DIF	.07	.09	.14	.13	.14	.19
		DAF and non-uniform DIF	.08	.09	.14	.10	.14	.19
	unmatched	No DIF and DAF	.08	.09	.14	.10	.12	.19
		DAF only	.08	.09	.13	.13	.14	.14
		Uniform DIF only	.07	.09	.14	.10	.13	.19
		Non-uniform DIF only	.07	.09	.14	.10	.13	.18
		DAF and uniform DIF	.08	.09	.13	.10	.13	.14
		DAF and non-uniform DIF	.08	.09	.13	.10	.13	.14

Table 4.3: RMSE of DAF Parameter γ over 25 Replications

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	No DIF and DAF	.15	.24	.25	.24	.29	.28
		DAF only	.17	.20	.30	.25	.28	.31
		Uniform DIF only	.12	.22	.26	.21	.27	.29
		Non-uniform DIF only	.17	.22	.25	.24	.26	.31
		DAF and uniform DIF	.14	.19	.32	.23	.22	.32
		DAF and non-uniform DIF	.15	.20	.31	.25	.24	.34
	unmatched	No DIF and DAF	.21	.19	.30	.22	.25	.37
		DAF only	.33	.33	.23	.37	.28	.37
		Uniform DIF only	.16	.19	.22	.24	.24	.39
		Non-uniform DIF only	.17	.21	.26	.22	.29	.37
		DAF and uniform DIF	.23	.22	.21	.35	.29	.35
		DAF and non-uniform DIF	.24	.24	.25	.37	.31	.37
1000/g	matched	No DIF and DAF	.18	.19	.21	.19	.15	.26
		DAF only	.14	.16	.29	.18	.19	.30
		Uniform DIF only	.13	.19	.21	.18	.15	.24
		Non-uniform DIF only	.14	.20	.22	.23	.15	.24
		DAF and uniform DIF	.13	.31	.28	.16	.21	.25
		DAF and non-uniform DIF	.22	.32	.31	.20	.20	.27
	unmatched	No DIF and DAF	.18	.15	.22	.24	.21	.26
		DAF only	.22	.19	.22	.17	.22	.36
		Uniform DIF only	.13	.17	.21	.25	.19	.25
		Non-uniform DIF only	.17	.17	.22	.19	.23	.26
		DAF and uniform DIF	.24	.21	.26	.28	.27	.36
		DAF and non-uniform DIF	.23	.22	.28	.32	.22	.30

Table 4.4: Summary of Means and Ranges of RMSEs of Attribute Level Parameters

			α	β	γ
Structure	simple	Mean	.32	.12	.22
		Range	(.04 -1.19)	(.07 -.20)	(.12 -.33)
	complex	Mean	.41	.15	.26
		Range	(.05 -1.32)	(.10 -.19)	(.15 -.39)
Attribute Discrimination	a=1	Mean	.08	.12	.21
		Range	(.04 -.13)	(.07 -.17)	(.12 -.37)
	a=2	Mean	.11	.13	.22
		Range	(.04 -.24)	(.09 -.16)	(.15 -.33)
	a=6	Mean	.91	.17	.28
		Range	(.57 -1.32)	(.12 -.20)	(.21 -.39)
Sample Size	500/g	Mean	.40	.15	.26
		Range	(.05 -1.32)	(.12 -.20)	(.12 -.39)
	1000/g	Mean	.34	.12	.22
		Range	(.04 -1.11)	(.07 -.19)	(.13 -.36)
Ability Distribution	matched	Mean	.33	.14	.22
		Range	(.04 -1.11)	(.07 -.20)	(.12 -.34)
	unmatched	Mean	.41	.14	.25
		Range	(.04 -1.32)	(.07 -.19)	(.13 -.39)
Scenarios	No DIF/DAF	Mean	.37	.14	.23
		Range	(.04 -1.20)	(.07 -.20)	(.15 -.37)
	DAF only	Mean	.37	.13	.25
		Range	(.06 -1.24)	(.07 -.20)	(.14 -.37)
	Uniform DIF only	Mean	.37	.14	.21
		Range	(.04 -1.29)	(.07 -.20)	(.12 -.39)
	Non-uniform DIF only	Mean	.37	.14	.22
		Range	(.05 -1.32)	(.07 -.20)	(.14 -.37)
	DAF and uniform DIF	Mean	.36	.13	.25
		Range	(.04 -1.22)	(.07 -.19)	(.13 -.36)
	DAF and non-uniform DIF	Mean	.37	.14	.27
		Range	(.04 -1.26)	(.08 -.19)	(.15 -.37)
Overall		Mean	.37	.14	.24
		Range	(.04 -1.32)	(.07-.20)	(.12 -.39)

4.1.2 The Type I Error Control of DAF

The Type I errors for DAF occur when an attribute is identified having DAF but DAF wasn't simulated for that attribute. In this study, each condition had 25 replications and each replication simulated 5 attributes. Under the conditions "No DAF and No DIF", "uniform DIF only" and "non-uniform DIF only" all five attributes were simulated as having no DAF. So the empirical Type I error rate was calculated as the percent of DAF detected out of 125 (= 25 replications × 5 attributes) no-DAF counts. Under the conditions of "DAF only", "Both DAF and uniform DIF" and "Both DAF and non-uniform DIF", 3 of 5 attributes were simulated to have no DAF. The empirical Type I error for these conditions was calculated as the percent of DAF detected out of 75 (= 25 replications × 3 attributes) no-DAF counts. Table 4.5 presents the empirical Type I error rates over all conditions.

To make meaningful power comparisons among conditions, Type I error should be controlled in each condition being compared. This is because lack of Type I error control is an indication that the model being simulated was not realized in the simulations. One result of this is that inflated Type I error rates result in overestimated power. Similarly, deflated Type I error rates result in underestimated power. The significance level was set as $\alpha = .05$ in this study. Thus, the empirical Type I error rate should be close to .05 in order to be considered as controlled. As a criterion for judging whether Type I error was controlled, Bradley's (1978) criterion was used. For a nominal level of .05, the Bradley criterion suggests a range of 025 to .075. Based on this criterion, 6 conditions yielded inflated Type I error rates and 56 conditions yielded deflated Type I error rates. According to table 4.5 there was no clear pattern about how the Type I error control of DAF increased or decreased across the conditions or which level of factors was more likely to yield inflated or deflated Type I error rates of DAF.

Presented in Table 4.6 are the marginal averages as well as the ranges of Type I errors under each level of each factor. It can be seen that the marginal Type I errors were all very close to .05 and varied less across different levels of each factor. Only two factors seemed to have some effect: The conditions with complex structure showed higher Type I error rates for DAF than those with simple structure; and Type I error rates for DAF tended be smaller as attribute discrimination increased. All other factors showed little difference among levels.

Table 4.5: Type I Error Rates for DAF over 25 Replications

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	No DIF and DAF	**.024-**	**.012-**	.040	.048	.072	**.024-**
		DAF only	**.013-**	.053	.053	.067	**.000-**	**.013-**
		Uniform DIF only	**.016-**	.032	.032	.064	**.080+**	**.024-**
		Non-uniform DIF only	**.024-**	.032	.048	.048	.072	.032
		DAF and uniform DIF	**.014-**	.040	.040	.040	**.000-**	**.014-**
		DAF and non-uniform DIF	.027	.029	.050	.067	**.000-**	**.013-**
	unmatched	No DIF and DAF	.040	.040	.048	**.000-**	**.016-**	**.024-**
		DAF only	.047	**.013-**	.013	.080	**.000-**	.027
		Uniform DIF only	**.018-**	.032	.045	.038	.040	**.024-**
		Non-uniform DIF only	**.024-**	.040	**.016-**	.040	.044	**.024-**
		DAF and uniform DIF	**.013-**	**.013-**	.044	.040	.032	.042
		DAF and non-uniform DIF	.013	.043	**.013-**	**.017-**	.043	.040
1000/g	matched	No DIF and DAF	.048	**.000-**	.032	.056	**.016-**	.032
		DAF only	.027	.027	.040	.027	.042	**.013-**
		Uniform DIF only	**.016-**	**.104+**	.032	.048	**.016-**	.032
		Non-uniform DIF only	**.024-**	**.088+**	.032	.048	**.008-**	**.024-**
		DAF and uniform DIF	**.013-**	.073	.027	.040	.053	**.013-**
		DAF and non-uniform DIF	.067	**.147+**	.067	**.013-**	.027	**.000-**
	unmatched	No DIF and DAF	.032	**.000-**	**.016-**	.040	**.024-**	.048
		DAF only	.027	**.013-**	.053	.056	**.013-**	.053
		Uniform DIF only	**.078+**	.046	**.024-**	.056	**.024-**	.048
		Non-uniform DIF only	.046	.050	**.024-**	.072	**.016-**	.046
		DAF and uniform DIF	.034	**.013-**	**.013-**	**.013-**	**.013-**	.053
		DAF and non-uniform DIF	.043	**.000-**	.026	**.013-**	.040	**.080+**

+ inflated Type I error; - deflated Type I error.

The number of conditions with uncontrolled Type I error rates for DAF was large, especially for the deflated cases. One possibility is that this occurred because only 25 replications were done for each condition and only 5 attributes were simulated in each replication. Regarding the number of no-DAF attributes (25 × 5 = 125 for No-DAF scenarios and 25 × 3 = 75 for DAF scenarios, small numbers of events (i.e., no-DAF attributes) will often

Table 4.6: Marginal Means and Ranges of Type I Error Rates for DAF

		Mean	Range
Structure	simple	.041	(.000-.147)
	complex	.046	(.000-.080)
Attribute	a=1	.046	(.000-.080)
Discrimination	a=2	.044	(.000-.147)
	a=6	.041	(.000-.080)
Sample Size	500/g	.043	(.000-.080)
	1000/g	.044	(.000-.147)
Ability	matched	.045	(.000-.147)
Distribution	unmatched	.044	(.000-.080)
Scenarios	No DIF/DAF	.045	(.000-.072)
	DAF only	.045	(.000-.080)
	Uniform DIF only	.044	(.000-.104)
	Non-uniform DIF only	.045	(.000-.088)
	DAF and uniform DIF	.043	(.000-.053)
	DAF and non-uniform DIF	.042	(.000-.147)
Overall		.044	(.000-.147)

result in a larger standard errors, accordingly the empirical Type I error will range more widely than conditions with larger numbers of replications. In addition, Type I errors for DIF-g and DIF-s reported in section 4.2.2 showed very good control. For both of these, there were more actual events included even though the number of replications was still 25. This is because each replication generated 25 items rather than 5 (i.e., the number of attributes). Second, the marginal means of Type I error rates in Table 4.6 were all very close to .05, since the number of trials dramatically increased when conditions were combined together. All of this suggests that the method in this study appeared to have reasonable control of Type I errors for DAF. Better control might be demonstrated if more replications were added in the future. (Given the complex model estimated using the current MCMC algorithm, however, estimation of another 25 replications for each condition is estimated to take an additional six to eight months.)

4.1.3 The Power of DAF

The percentage of times DAF was identified, when DAF was simulated, was calculated as the estimate of the empirical power. In this study, the conditions "DAF only", "Both DAF and uniform DIF", and "Both DAF and non-uniform DIF" included simulations of DAF in 2 of the 5 attributes in each replication. Table 4.7 only provided the empirical power, therefore, for these three scenarios. Recall, the Type I errors of DAF under many conditions in section 4.1.2 were either inflated or deflated, and it's not clear whether they were truly out of control or appeared so due to limited numbers of replications.

As can be seen in Table 4.7, it appeared that those conditions with larger sample size, simple structure, lower attribute discrimination (i.e., $a = 1$ or $a = 2$), and matched general ability distribution had higher power than the conditions with smaller sample size, complex structure, higher attribute discrimination (i.e., $a = 6$), and unmatched general ability distribution.

Table 4.8 provides a somewhat clearer picture about the effect of each factor on the power of DAF by showing the marginal means of power percentages: Complex structure had 17% lower power rates of detecting DAF than simple structure; the power was almost the same when attribute discrimination equaled 1 or 2 (ranging from .95 to .96), but dropped to .76 when attribute discrimination was 6; the average power rate increased .08 from sample size 500/group to 1000/group, decreased .10 from matched ability distribution to unmatched ability distribution, and had no obvious change among three of the condition scenarios.

Table 4.7: Power of DAF over 25 Replications

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	DAF only	.98	1.00	.98	.98	1.00	.66
		DAF and uniform DIF	1.00	1.00	.98	1.00	.98	.60
		DAF and non-uniform DIF	1.00	1.00	.94	1.00	.96	.60
	unmatched	DAF only	.98	1.00	.82	.88	.72	.36
		DAF and uniform DIF	.96	1.00	.72	.86	.66	.47
		DAF and non-uniform DIF	1.00	1.00	.74	.68	.68	.34
1000/g	matched	DAF only	1.00	1.00	1.00	1.00	1.00	.74
		DAF and uniform DIF	1.00	1.00	1.00	1.00	1.00	.78
		DAF and non-uniform DIF	1.00	1.00	1.00	1.00	1.00	.72
	unmatched	DAF only	1.00	1.00	.98	.92	.94	.66
		DAF and uniform DIF	1.00	1.00	.98	.94	.94	.62
		DAF and non-uniform DIF	1.00	1.00	.94	.82	.84	.62

Table 4.8: Marginal Means and Ranges of Empirical Power of DAF

		Mean	Range
Structure	simple	.97	(.72-1.00)
	complex	.80	(.34-1.00)
Attribute Discrimination	a=1	.96	(.68-1.00)
	a=2	.95	(.66-1.00)
	a=6	.76	(.34-1.00)
Sample Size	500/g	.85	(.34-1.00)
	1000/g	.93	(.62-1.00)
Ability Distribution	matched	.94	(.60-1.00)
	unmatched	.84	(.34-1.00)
Scenarios	DAF only	.90	(.36-1.00)
	DAF and uniform DIF	.90	(.47-1.00)
	DAF and non-uniform DIF	.87	(.34-1.00)
Overall		.89	(.34-1.00)

In Table 4.7, for simple structure, all conditions except the conditions with unmatched ability distribution and attribute discrimination $a = 6$ had power close to 1. Under complex structure, in general, the power was much lower, but some conditions with matched ability distribution and attribute discrimination $a = 1$ or $a = 2$ still yielded power as high as 1. This suggested one or more "good situations" could compensate certain "bad situations." Figures 4.1 to 4.3 indicated that simple structure generally had higher power than complex structure. In addition, power for simple structure didn't decline as much as that for complex structure when attribute discrimination was 6, sample size was 500/group, or ability distribution was unmatched. Similarly, Figures 4.1.3.4 to 4.1.3.5 demonstrated that larger sample size was more powerful than small sample size. Moreover, the power rates for large sample size don't show a decline like those under small sample size, when attribute discrimination was 6 or ability distribution was unmatched.

Figure 4.1: The Interaction between Q-matrix Structure and Attribute Discrimination

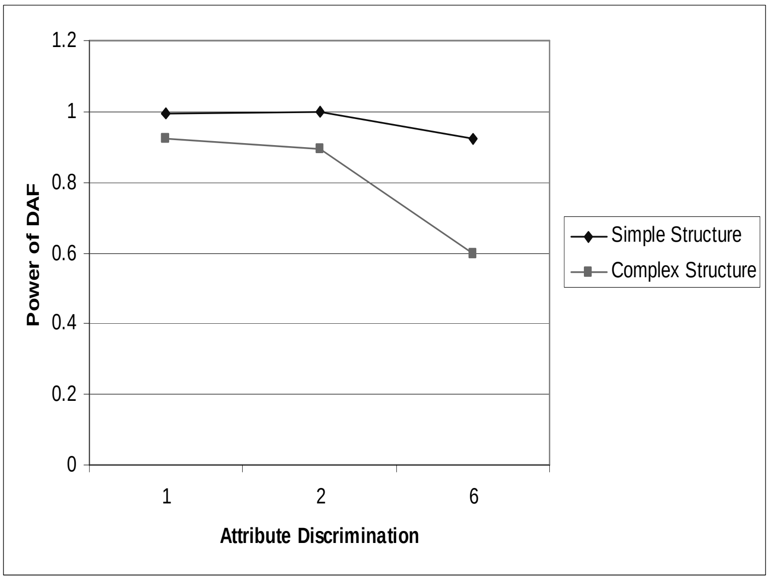


Figure 4.2: The Interaction between Q-matrix Structure and Sample Size

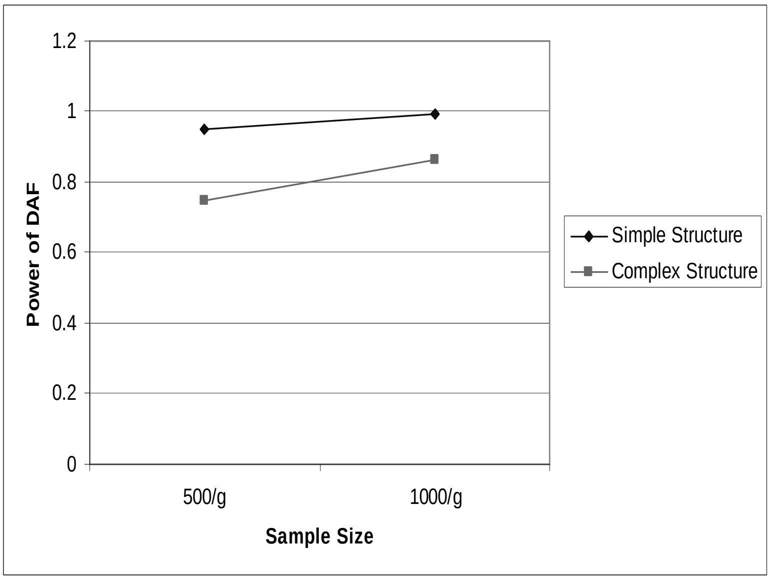


Figure 4.3: The Interaction between Q-matrix Structure and Ability Distribution

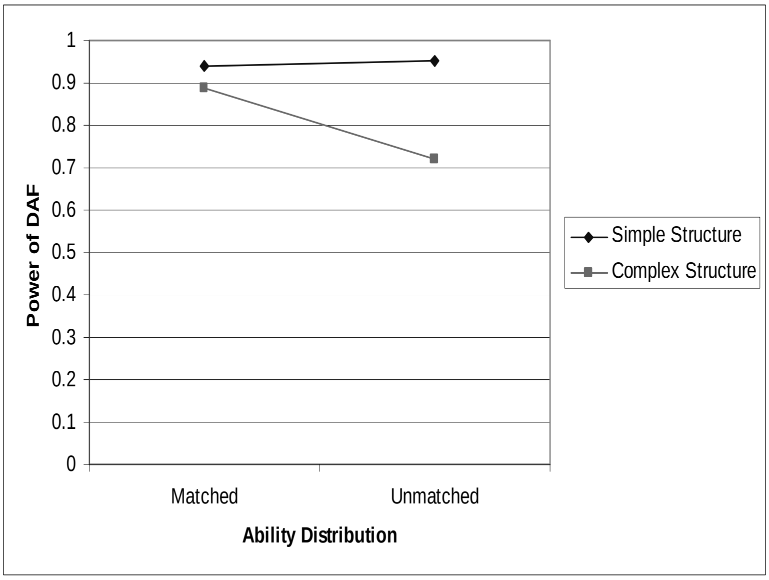


Figure 4.4: The Interaction between Sample Size and Attribute Discrimination

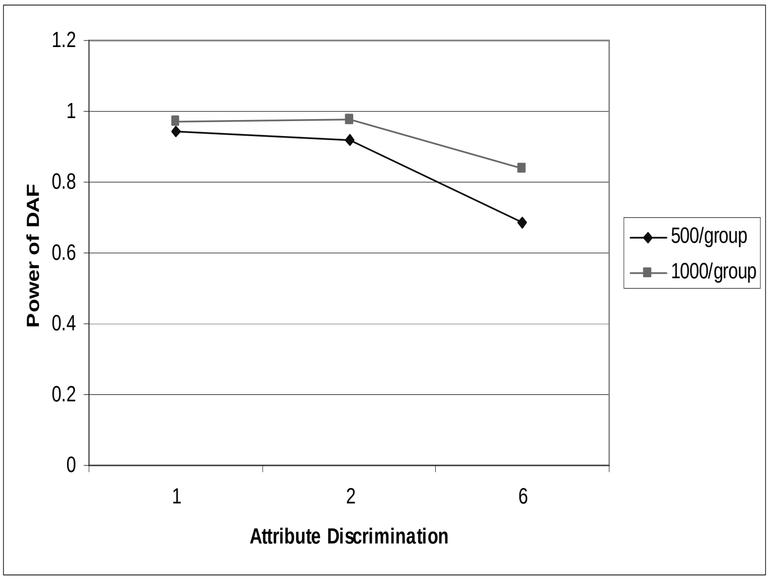
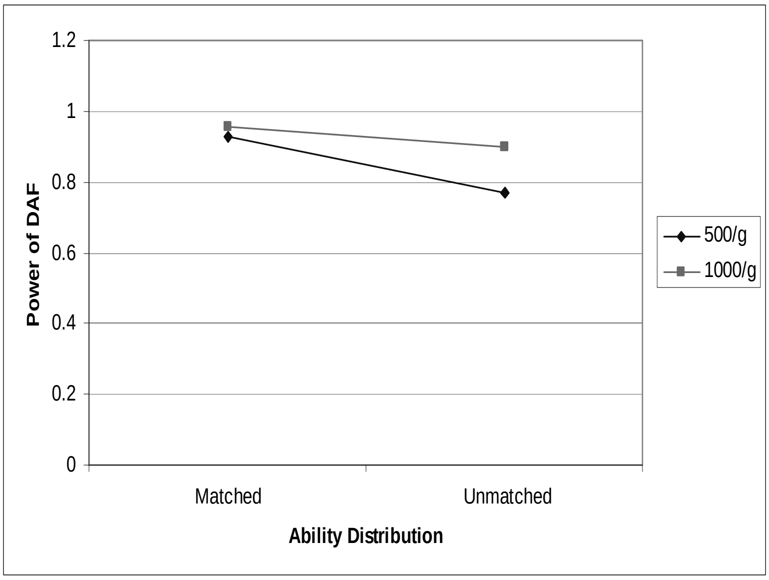


Figure 4.5: The Interaction between Sample Size and Ability Distribution



4.2 DIF Detection

4.2.1 Recovery of Lower-Level Parameters

Lower-level item parameters in the modified higher-order DINA include the “guessing” parameter, g, “slipping” parameter, s, and examinee attribute mastery parameter, α. Table 4.9 provide the correct classification rates for α. Tables 4.10 and 4.11 provided RMSEs for g and s under each simulated condition, since DIF-g and DIF-s essentially reflect group differences in parameter g and s. Table 4.12 presents the mean and range of RMSEs of g, s and correct classification rates of α under each level of manipulated factors.

In Table 4.9, it can be seen that the correct classification rates of attribute mastery were relatively high for simple structure conditions, ranging from .91 to .97. They decreased for the complex structure conditions except for those conditions for which $a = 6$. The main effects of the Q-matrix structure and attribute discrimination are also evident in Table 4.12. For the DINA model, the probability of a correct response to item i equals g_i whether examinees miss one or more attributes required by item i. Under simple structure, when an examinee misses the item not due to slipping, it means that the single attribute required by that item is not mastered. With complex structure, however, it can be difficult to distinguish which particular attribute(s) an examinee has not mastered, if the item is missed due to other than slipping. For this reason, a test constructed with simple structure will tend to classify examinees more correctly, if the number of items measuring each attribute is large enough. The relationship between correct classification and attribute discrimination is straightforward: The attribute mastery probability will be more sensitive to the general ability level, when it has higher discrimination. In such a case, small differences in general ability can result in large differences in attribute mastery. Therefore, the attribute with higher discrimination is more informative and more able to distinguish between masters and non-masters. The larger sample size did not improve estimation of α. This is similar to results from IRT in which the number of examinees tends to improve estimation of item parameters but not examinee parameters.

Tables 4.10 and 4.11 provide the RMSEs for g and s under each simulated condition. Both parameters were recovered well. The RMSEs of g ranged from .02 to .05 (.02 for most of conditions), and for s they ranged from .02 to .06. The recovery of g was slightly better than the recovery of s.

The combination of simple structure, $a = 1$, 500/group, and matched ability in Table 4.10 had the highest RMSE for g. The combination of complex structure, $a = 1$, 500/group, and unmatched ability in Table 4.11 had the highest RMSE for s. This was consistent with the recovery results for g and s with respect to sample size and attribute discrimination. It is clear that recovery in the smaller sample size was less accurate. It is interesting to note that recovery was poorer for lower attribute discrimination than for higher attribute discrimination conditions. One possible explanation is that classification may be less accurate when attribute discrimination is lower, thereby resulting in less accurate recovery in g and s.

Higher RMSEs for g were observed for both the simple structure and matched ability conditions. This was contrary to the pattern observed for RMSEs of s, but is consistent with the definitions of g and s. Recall that "g is the probability of a correct response for someone classified as lacking at least one required attribute" (p. 11) and "s is the probability of missing the item for someone who is classified as mastering all required attributes" (p. 10). Estimation of g is conditioned on the status of non-masters, and estimation of s is conditioned on the status of masters. What this means is that the estimation of g actually relies on the sample size of non-masters, and the estimation of s relies on the sample size of masters. If the total sample size is not large enough (e.g., 500/group), more masters in the sample means there will be too few non-masters and, conversely, more non-masters in the sample means too few masters. In the simple structure conditions, each question only required a single attribute, so more examinees in the sample were masters of that item than for the same sample size in the complex structure condition. This situation advantaged estimation of the slipping parameter, s, but actually disadvantaged estimation of the guessing parameter, g. This may explain why g had higher RMSEs than s in the simple structure conditions, but s had higher RMSEs than g in the complex structure conditions.

One possible explanation for the poorer recovery results for g in the matched ability distribution might be due to the design of the simulation study. In the matched ability condition abilities were generated using $N(0, 1)$ for both reference group and focal group, whereas in the unmatched ability distribution condition, abilities were generated based on $N(0, 1)$ for the reference group and $N(-1, 1)$ for the focal group. This resulted in more high ability examinees in the matched ability distribution condition, and therefore, more masters, since the mastery status of each attribute was dependent on general ability (see equation 2.5 in Chapter 2). For the same reason, the more masters there were, the fewer non-masters there were in the sample. This situation improved estimation of the slipping parameter but not the guessing parameter. Consequently, in the matched ability condition, recovery was poorer for g but better for s.

In Table 4.12, the pattern of recovery was similar to that shown in Tables 4.10 and 4.11. These scenarios appear to have had no impact on estimation of either g or s; the effect of attribute discrimination and sample size had the same pattern of recovery on g and s; the effect of Q-matrix structure and ability distribution was reversed on g and s.

Table 4.9: Mean Proportions of Correct Classification Rates, α, over 25 Replications

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	No DIF and DAF	.92	.93	.96	.83	.86	.94
		DAF only	.92	.93	.96	.82	.86	.93
		Uniform DIF only	.92	.93	.96	.83	.87	.94
		Non-uniform DIF only	.91	.92	.96	.82	.86	.94
		DAF and uniform DIF	.92	.93	.94	.83	.86	.93
		DAF and non-uniform DIF	.91	.93	.96	.82	.85	.93
	unmatched	No DIF and DAF	.92	.94	.96	.82	.87	.94
		DAF only	.92	.94	.97	.81	.86	.94
		Uniform DIF only	.92	.94	.97	.82	.87	.95
		Non-uniform DIF only	.92	.93	.96	.82	.87	.94
		DAF and uniform DIF	.93	.94	.97	.82	.86	.94
		DAF and non-uniform DIF	.92	.93	.96	.81	.86	.94
1000/g	matched	No DIF and DAF	.92	.93	.96	.83	.87	.94
		DAF only	.92	.93	.96	.83	.86	.93
		Uniform DIF only	.92	.93	.96	.83	.87	.94
		Non-uniform DIF only	.92	.92	.96	.83	.87	.93
		DAF and uniform DIF	.92	.93	.96	.83	.87	.93
		DAF and non-uniform DIF	.92	.93	.96	.82	.86	.93
	unmatched	No DIF and DAF	.93	.94	.97	.83	.87	.95
		DAF only	.92	.94	.97	.82	.87	.94
		Uniform DIF only	.92	.94	.97	.82	.87	.95
		Non-uniform DIF only	.92	.93	.96	.82	.87	.94
		DAF and uniform DIF	.93	.94	.97	.82	.87	.94
		DAF and non-uniform DIF	.92	.93	.96	.81	.87	.94

Table 4.10: RMSEs of Guessing Parameter, g, over 25 Replications

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	No DIF and DAF	.03	.03	.03	.02	.03	.03
		DAF only	.04	.04	.03	.02	.02	.03
		Uniform DIF only	.04	.03	.03	.03	.03	.03
		Non-uniform DIF only	.04	.03	.03	.03	.03	.03
		DAF and uniform DIF	.03	.03	.04	.02	.03	.03
		DAF and non-uniform DIF	.04	.03	.03	.02	.03	.03
	unmatched	No DIF and DAF	.04	.03	.03	.02	.02	.02
		DAF only	.05	.03	.03	.02	.02	.02
		Uniform DIF only	.03	.03	.03	.02	.02	.02
		Non-uniform DIF only	.03	.03	.03	.02	.02	.02
		DAF and uniform DIF	.03	.03	.03	.02	.02	.02
		DAF and non-uniform DIF	.03	.03	.03	.02	.02	.02
1000/g	matched	No DIF and DAF	.02	.02	.02	.02	.02	.02
		DAF only	.02	.02	.02	.02	.02	.02
		Uniform DIF only	.02	.02	.02	.02	.02	.02
		Non-uniform DIF only	.02	.02	.02	.02	.02	.02
		DAF and uniform DIF	.02	.02	.02	.02	.02	.02
		DAF and non-uniform DIF	.02	.02	.02	.02	.02	.02
	unmatched	No DIF and DAF	.02	.02	.02	.02	.02	.02
		DAF only	.03	.02	.02	.02	.02	.02
		Uniform DIF only	.02	.02	.02	.02	.02	.02
		Non-uniform DIF only	.02	.02	.02	.02	.02	.02
		DAF and uniform DIF	.02	.02	.02	.02	.02	.02
		DAF and non-uniform DIF	.02	.02	.02	.02	.02	.02

Table 4.11: RMSEs of Slipping Parameter, s, over 25 Replications

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	No DIF and DAF	.03	.03	.03	.04	.03	.03
		DAF only	.04	.04	.03	.04	.03	.03
		Uniform DIF only	.04	.03	.03	.04	.03	.03
		Non-uniform DIF only	.04	.03	.03	.04	.03	.03
		DAF and uniform DIF	.03	.03	.03	.04	.03	.03
		DAF and non-uniform DIF	.04	.03	.03	.04	.03	.03
	unmatched	No DIF and DAF	.04	.04	.04	.06	.05	.04
		DAF only	.05	.04	.04	.06	.05	.04
		Uniform DIF only	.04	.04	.04	.05	.05	.04
		Non-uniform DIF only	.04	.04	.04	.06	.05	.04
		DAF and uniform DIF	.03	.04	.04	.06	.05	.04
		DAF and non-uniform DIF	.04	.04	.04	.06	.05	.04
1000/g	matched	No DIF and DAF	.02	.02	.02	.03	.03	.02
		DAF only	.02	.02	.02	.03	.02	.02
		Uniform DIF only	.02	.02	.02	.03	.03	.03
		Non-uniform DIF only	.02	.02	.02	.04	.03	.02
		DAF and uniform DIF	.02	.02	.02	.03	.03	.02
		DAF and non-uniform DIF	.02	.02	.02	.03	.03	.02
	unmatched	No DIF and DAF	.03	.03	.03	.04	.04	.03
		DAF only	.04	.03	.03	.04	.04	.03
		Uniform DIF only	.03	.03	.05	.05	.04	.03
		Non-uniform DIF only	.03	.03	.03	.03	.04	.03
		DAF and uniform DIF	.03	.03	.03	.04	.04	.03
		DAF and non-uniform DIF	.03	.03	.03	.05	.04	.03

Table 4.12: The Marginal Means and Ranges of RMSEs for Item Level Parameters

			g	s	α
Structure	simple	Mean	.026	.030	.94
		Range	(.02 -.05)	(.02 - .05)	(.91 -.97)
	complex	Mean	.020	.037	.88
		Range	(.02 -.03)	(.02 - .06)	(.81 -.95)
Attribute Discrimination	a=1	Mean	.025	.03	.87
		Range	(.02 -.05)	(.02 - .06)	(.81 -.93)
	a=2	Mean	.023	.033	.90
		Range	(.02 -.04)	(.02 - .05)	(.85 -.94)
	a=6	Mean	.022	.030	.95
		Range	(.02 -.04)	(.02 - .04)	(.93 -.97)
Sample Size	500/g	Mean	.028	.038	.91
		Range	(.02 -.05)	(.03 - .06)	(.81 -.97)
	1000/g	Mean	.019	.028	.91
		Range	(.02 -.03)	(.02 - .04)	(.81 -.97)
Ability Distribution	matched	Mean	.024	.028	.91
		Range	(.02 -.04)	(.02 - .04)	(.82 -.96)
	unmatched	Mean	.022	.039	.91
		Range	(.02 -.05)	(.02 - .06)	(.81 -.97)
Scenarios	No DIF/DAF	Mean	.023	.033	.91
		Range	(.02 -.04)	(.02 - .06)	(.82 -.97)
	DAF only	Mean	.024	.034	.91
		Range	(.02 -.05)	(.02 - .06)	(.81 -.97)
	Uniform DIF only	Mean	.023	.032	.91
		Range	(.02 -.04)	(.02 - .06)	(.82 -.97)
	Non-uniform DIF only	Mean	.023	.034	.91
		Range	(.02 -.04)	(.02 - .06)	(.81 -.96)
	DAF and uniform DIF	Mean	.023	.032	.91
		Range	(.02 -.04)	(.02 - .06)	(.82 -.97)
	DAF and non-uniform DIF	Mean	.023	.034	.90
		Range	(.02 -.04)	(.02 - .06)	(.81 -.96)
Overall		Mean	.023	.033	.91
		Range	(.02 -.05)	(.02 - .06)	(.81-.97)

4.2.2 The Type I error of DIF-g and DIF-s

The Type I errors for DIF-g and DIF-s occur when an item is identified having DIF-g or DIF-s, but either DIF-g or DIF-s were not simulated, respectively. In this study, each condition had 25 replications with 25 items for each replication. Under the conditions of "No DAF and DIF" and "DAF only," all 25 items were simulated with either no DIF-g or no DIF-s. The empirical Type I error rate for DIF-g was calculated as the percent of DIF-g detected out of 625 (= 25 replications × 25 items) no-DIF counts. The same was done for the empirical Type I error of DIF-s. Under the conditions of "uniform DIF only", "non-uniform DIF only", "Both DAF and uniform DIF" and "Both DAF and non-uniform DIF", 20 of 25 items were simulated with either no DIF-g or no DIF-s. Therefore, the empirical Type I error rate for DIF-g was calculated as the percent of DIF-g detected out of 500 (= 25 replications × 20 items) no-DIF counts. The same was done for DIF-s. Table 4.13 and Table 4.14 presented the empirical Type I error rates of DIF-g and DIF-s over all conditions, respectively.

As was noted for Type I error for DAF, the Type I error for DIF-g or for DIF-s must be controlled in order to estimate the power of DIF-g or DIF-s. The nominal level for these Type I error rates was set at $\alpha = .05$. As was the case for DAF, the range of accepted Type I error rate followed Bradley's (1978) criterion of .025 to .075. Results in Tables 4.13 and 4.14 indicated that the empirical Type I error rate was controlled in all conditions: Type I error rate for DIF-g ranged from .028 to .062 and for DIF-s from .026 to .68. Consistent with the Type I error results for DAF, Type I error rates for DIF-g and DIF-s did not show a clear pattern regarding which conditions had lower or higher rates of Type I errors. Marginal means and ranges of Type I error rates for DIF-g and DIF-s were reported at each level of five factors in Table 4.15. Type I errors for DIF-g was consistent among different levels of all five factors. Type I errors of DIF-s appeared to be sensitive to sample size and attribute discrimination. The Type I errors for DIF-s increased to .053 from .041 when the sample size increased to 1000 examinees per group from 500 examinees per group. In addition, higher attribute discrimination (i.e., $a = 2$ or $a = 6$) yielded relatively higher Type I error for DIF-s than lower attribute discrimination (i.e., $a = 1$).

Table 4.13: Type I Error of DIF-g over 25 Replications

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	No DIF and DAF	.046	.040	.046	.045	.043	.056
		DAF only	.046	.056	.045	.043	.043	.051
		Uniform DIF only	.052	.034	.044	.046	.038	.045
		Non-uniform DIF only	.050	.062	.042	.044	.040	.050
		DAF and uniform DIF	.042	.048	.038	.044	.038	.045
		DAF and non-uniform DIF	.044	.039	.044	.034	.042	.046
	unmatched	No DIF and DAF	.048	.038	.051	.042	.032	.043
		DAF only	.054	.045	.045	.046	.040	.042
		Uniform DIF only	.039	.038	.044	.034	.034	.044
		Non-uniform DIF only	.044	.032	.042	.034	.034	.044
		DAF and uniform DIF	.045	.040	.047	.046	.038	.047
		DAF and non-uniform DIF	.044	.038	.050	.042	.044	.044
1000/g	matched	No DIF and DAF	.035	.040	.034	.046	.042	.045
		DAF only	.045	.045	.038	.034	.033	.040
		Uniform DIF only	.048	.038	.040	.048	.050	.042
		Non-uniform DIF only	.040	.032	.038	.042	.048	.044
		DAF and uniform DIF	.042	.028	.042	.040	.044	.048
		DAF and non-uniform DIF	.042	.030	.038	.050	.038	.052
	unmatched	No DIF and DAF	.037	.043	.037	.040	.048	.037
		DAF only	.048	.043	.034	.042	.043	.043
		Uniform DIF only	.046	.046	.030	.044	.040	.036
		Non-uniform DIF only	.042	.048	.036	.050	.034	.036
		DAF and uniform DIF	.054	.046	.028	.034	.046	.042
		DAF and non-uniform DIF	.032	.044	.030	.040	.048	.044

Table 4.14: Type I Error of DIF-s over 25 Replications

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	No DIF and DAF	.035	.042	.053	.035	.037	.035
		DAF only	.038	.050	.045	.045	.043	.038
		Uniform DIF only	.034	.046	.042	.034	.042	.036
		Non-uniform DIF only	.040	.050	.048	.026	.048	.040
		DAF and uniform DIF	.028	.050	.038	.034	.046	.038
		DAF and non-uniform DIF	.034	.039	.046	.030	.048	.042
	unmatched	No DIF and DAF	.042	.038	.062	.045	.048	.045
		DAF only	.061	.034	.051	.027	.042	.035
		Uniform DIF only	.026	.048	.060	.038	.046	.046
		Non-uniform DIF only	.036	.040	.062	.038	.048	.046
		DAF and uniform DIF	.027	.040	.052	.026	.062	.038
		DAF and non-uniform DIF	.026	.028	.060	.026	.026	.034
1000/g	matched	No DIF and DAF	.045	.046	.043	.054	.062	.043
		DAF only	.050	.048	.040	.042	.038	.050
		Uniform DIF only	.048	.060	.054	.068	.066	.046
		Non-uniform DIF only	.046	.068	.048	.058	.062	.050
		DAF and uniform DIF	.048	.054	.050	.050	.068	.054
		DAF and non-uniform DIF	.058	.066	.050	.046	.066	.052
	unmatched	No DIF and DAF	.046	.051	.056	.054	.062	.058
		DAF only	.040	.053	.058	.067	.061	.045
		Uniform DIF only	.048	.056	.066	.062	.056	.044
		Non-uniform DIF only	.052	.048	.064	.062	.054	.064
		DAF and uniform DIF	.038	.050	.056	.052	.060	.032
		DAF and non-uniform DIF	.052	.048	.060	.058	.056	.042

Table 4.15: Marginal Means and Ranges of Type I Errors of DIF-g and DIF-s

		DIF-g		DIF-s	
		Mean	Range	Mean	Range
Structure	simple	.042	(.028-.062)	.047	(.026-.068)
	complex	.042	(.032-.056)	.047	(.026-.068)
Attribute Discrimination	a=1	.043	(.034-.054)	.043	(.026-.068)
	a=2	.041	(.028-.062)	.050	(.026-.068)
	a=6	.042	(.028-.056)	.048	(.026-.064)
Sample Size	500/g	.043	(.032-.062)	.041	(.026-.062)
	1000/g	.041	(.028-.052)	.053	(.032-.068)
Ability Distribution	matched	.043	(.028-.062)	.047	(.026-.068)
	unmatched	.041	(.028-.054)	.048	(.026-.067)
Scenarios	No DIF/DAF	.042	(.032-.056)	.047	(.035-.062)
	DAF only	.044	(.033-.056)	.046	(.028-.067)
	Uniform DIF only	.042	(.030-.052)	.049	(.026-.066)
	Non-uniform DIF only	.042	(.032-.062)	.050	(.026-.068)
	DAF and uniform DIF	.042	(.028-.054)	.045	(.027-.066)
	DAF and non-uniform DIF	.042	(.030-.052)	.046	(.035-.062)
Overall		.042	(.028-.062)	.047	(.026-.068)

4.2.3 The Power of DIF-g and DIF-s

The percentage of identified DIF-g or DIF-s, when DIF-g or DIF-s, respectively, was simulated was calculated as the estimate of the percent of correct detections for DIF-g or DIF-s. Four scenarios were examined: "uniform DIF only", "non-uniform DIF only", "Both DAF and uniform DIF" and "Both DAF and non-uniform DIF". For each of these conditions, five of 25 items were simulated with both DIF-g and DIF-s in the same direction or different direction for each replication.

Table 4.16 and Table 4.17 provided estimates of empirical power of these four conditions for DIF-g and DIF-s, respectively. The power of DIF-g and DIF-s varied across the simulation conditions: The power of DIF-g ranged from .55 to 1.00 and the power of DIF-s ranged from .09 to 1.00. Consistent with the pattern of the recovery of g and s discussed in the Section 4.2.1, the block with simple structure, $a = 1$, 500 examinees per group, and matched ability (see Table 4.16) had the lowest power of DIF-g. This same block had the highest RMSEs for g (see Table 4.10). The block with complex structure, $a = 1$, 500 examinees per group, and unmatched ability (see Table 4.17) had the lowest power of DIF-s. This was the same block that had the highest RMSEs for s (see Table 4.11). That is, the block with highest RMSEs for g and s had the lowest power for DIF-g and DIF-s. To understand this result, it is important to recall that both DIF-g or DIF-s were detected based on the amounts of group difference in the g parameters or the s parameters, respectively.

As can be seen in Tables 4.16 and 4.17, the conditions with the larger sample size and higher attribute discrimination had higher power than the conditions with the smaller sample size and lower attribute discrimination. Table 4.18 presents marginal means and ranges of power for both DIF-g and DIF-s. The effect of the Q-matrix structure and matched vs unmatched ability distributions was reversed for DIF-g and DIF-s: Complex structure and unmatched ability produced higher power for DIF-g; simple structure and matched ability produced higher power for DIF-s. These power patterns were similar to the patterns of RMSEs for DIF-g and DIF-s, respectively.

The power rates of DIF-g and DIF-s appeared to essentially be related to the sample size of non-masters and masters. The conditions with the large sample size, complex structure and unmatched ability distribution produced more non-masters, and the conditions with the large sample size, simple structure and matched ability distribution produced more masters (as explained in Section 4.2.1). The effect of having more non-masters is that DIF-g is estimated more precisely. This, in turn, results in higher power of DIF-g. Likewise, more masters results in more precise estimation of DIF-s, thereby resulting in higher power of DIF-s.

As can be seen in Table 4.18, the power of DIF-g did not vary greatly across scenarios, although the power of DIF-s did. That is, the scenarios with non-uniform DIF (i.e., "non-uniform DIF only" and "DAF and non-uniform DIF") clearly had lower power of DIF-s than the scenarios with uniform DIF (i.e., "uniform DIF only" and "DAF and uniform DIF"), even though RMSEs of DIF-s showed very small differences across these scenarios (see Table 4.12). This result may possibly be related to the design of the uniform and non-uniform DIF conditions in the simulation study. In the simulation study, "five items were generated by increasing guessing parameters by .1 and decreasing the slip parameters by .1 for the focal group relative to the reference group" for uniform DIF (p. 33) and "the slip and guessing parameters were simulated in the five DIF items by increasing these parameters by .1 in the focal group" for non-uniform DIF (p. 34). For both uniform DIF and non-uniform DIF, the guessing g parameter was simulated to increase by .1 in the focal group. That is, the change for g was the same for different DIF scenarios. However, the slipping parameter, s, was decreased by .1 for uniform DIF and increased by .1 for non-uniform DIF in the focal group. Thus, uniform DIF scenarios had lower slipping parameters and non-uniform DIF scenarios had higher slipping parameters in the focal group. The slipping parameter is a kind of error parameter for detecting masters. As expected, lower slipping parameters in the uniform DIF scenarios resulted in more correct classifications of masters and higher slipping parameters in non-uniform DIF scenarios resulted in fewer correct classifications for masters.

Consequently, the power of DIF-s for uniform DIF scenarios was higher than the power of DIF-s for non-uniform DIF scenarios.

Table 4.16: Power of DIF-g over 25 Replications

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	Uniform DIF only	**.55**	.58	.71	.74	.74	.74
		Non-uniform DIF only	**.58**	.58	.72	.78	.77	.75
		DAF and uniform DIF	**.56**	.61	.64	.80	.78	.78
		DAF and non-uniform DIF	**.60**	.63	.70	.82	.79	.78
	unmatched	Uniform DIF only	.65	.72	.82	.82	.89	.86
		Non-uniform DIF only	.69	.71	.81	.84	.88	.89
		DAF and uniform DIF	.65	.70	.82	.83	.80	.88
		DAF and non-uniform DIF	.66	.73	.80	.82	.83	.90
1000/g	matched	Uniform DIF only	.78	.90	.91	.99	.99	.92
		Non-uniform DIF only	.79	.90	.94	1.00	.99	.96
		DAF and uniform DIF	.72	.86	.91	.96	.99	.96
		DAF and non-uniform DIF	.86	.87	.94	.97	.98	.98
	unmatched	Uniform DIF only	.88	.98	.98	1.00	1.00	1.00
		Non-uniform DIF only	.89	.98	.99	.98	1.00	1.00
		DAF and uniform DIF	.91	.98	.99	.98	.98	.94
		DAF and non-uniform DIF	.90	.97	.99	.98	.98	.99

Table 4.17: Power of DIF-s over 25 Replications

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	Uniform DIF only	.74	.79	.86	.50	.72	.86
		Non-uniform DIF only	.55	.54	.66	.34	.60	.65
		DAF and uniform DIF	.70	.75	.80	.43	.70	.83
		DAF and non-uniform DIF	.49	.59	.67	.31	.50	.62
	unmatched	Uniform DIF only	.64	.58	.51	**.27**	.30	.43
		Non-uniform DIF only	.37	.30	.30	**.14**	.17	.33
		DAF and uniform DIF	.63	.55	.51	**.15**	.23	.43
		DAF and non-uniform DIF	.35	.32	.30	**.09**	.15	.31
1000/g	matched	Uniform DIF only	.94	.95	.98	.70	.90	.93
		Non-uniform DIF only	.85	.89	.94	.43	.83	.94
		DAF and uniform DIF	.91	.94	.98	.72	.88	.95
		DAF and non-uniform DIF	.90	.89	.91	.65	.79	.92
	unmatched	Uniform DIF only	.83	.78	.81	.38	.54	.72
		Non-uniform DIF only	.66	.67	.66	.64	.40	.56
		DAF and uniform DIF	.86	.76	.83	.40	.47	.70
		DAF and non-uniform DIF	.66	.66	.66	.27	.37	.62

Table 4.18: Marginal Means and Ranges of Power Rates of DIF-g and DIF-s

		DIF-*g*		DIF-*s*	
		Mean	Range	Mean	Range
Structure	simple	.79	(.55-.99)	.70	(.49-.98)
	complex	.90	(.74-1.00)	.54	(.09-.90)
Attribute Discrimination	a=1	.81	(.55-1.00)	.55	(.09-.94)
	a=2	.85	(.58-1.00)	.62	(.15-.95)
	a=6	.88	(.64-1.00)	.69	(.30-.98)
Sample Size	500/g	.74	(.55-.90)	.49	(.09-.86)
	1000/g	.94	(.72-1.00)	.74	(.27-.98)
Ability Distribution	matched	.81	(.55-1.00)	.75	(.31-.98)
	unmatched	.88	(.65-1.00)	.48	(.09-.86)
Scenarios	Uniform DIF only	.84	(.55-1.00)	.69	(.27-.98)
	Non-uniform DIF only	.85	(.58-1.00)	.56	(.14-.94)
	DAF and uniform DIF	.83	(.56-.99)	.67	(.43-.98)
	DAF and non-uniform DIF	.85	(.60-.99)	.54	(.09-.92)
Overall		.84	(.55-1.00)	.62	(.09-.98)

In Tables 4.16 and 4.17, the block with the lowest power for DIF-g or DIF-s was shown in bold, but there were several blocks in tables indicating the same high level of power (ranging from .90 to 1.00). That means the higher power was reached not just by the combination of all good situations, but also by the combination of some good situations with some bad situations. For example, when the conditions meet 1000 examinees per group and complex structure, the power rates for DIF-g were all close to 1 even for $a = 1$ or for matched ability distribution. This would be a bad situation, in other words, for the power of DIF-g. This finding was similar to that for the power of DAF. That is, combinations including some bad situations (i.e., situations in which power of DAF was poor) can be compensated by good situations (i.e., ones which favor DAF detection). This suggested that some interactions existed between the factors manipulated in this study. The presence of these interactions, in fact, can be seen in Figures 4.6 to 4.8.

Figures 4.6 to 4.8 show the differences in power of DIF-g between simple structure and complex structure were reduced when attribute discrimination increased to 6, when sample size increased to 1000 examinees per group, or when ability distribution was unmatched, respectively. Figure 4.9 indicates the difference in the power of DIF-g between small sample size and large sample size was the same when attribute discrimination $a = 1$ and $a = 2$, but decreased when $a = 6$. In conditions of attribute discrimination $a = 6$, large sample size and unmatched ability distribution, or complex structure, power was greater than in the opposite condition for DIF-g. As an example, even though the power of DIF-g was higher for complex structure than for simple structure, the difference was smaller when this condition was combined with one of the other conditions for which power was also good (e.g., with high attribute discrimination, larger sample size, or unmatched distribution).

Figure 4.6: The Interaction between Attribute Discrimination and Q-matrix Structure on the Power of DIF-g

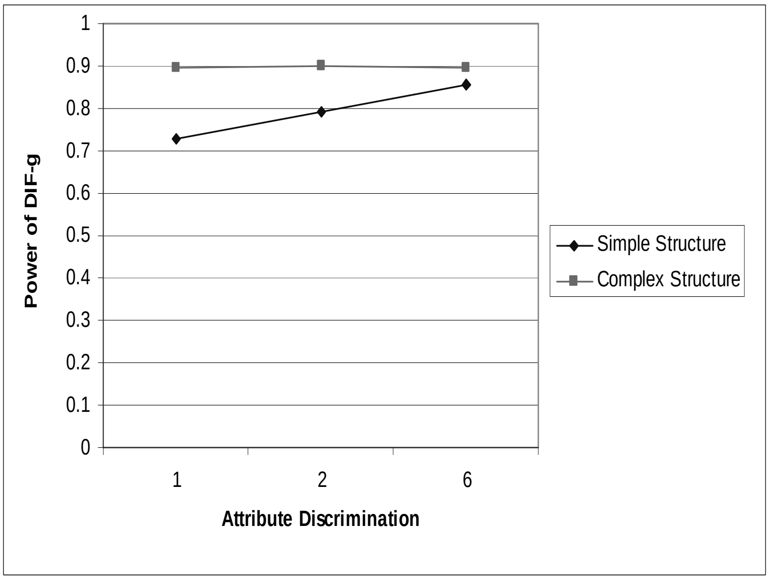


Figure 4.7: The Interaction between Sample Size and Q-matrix Structure on the Power of DIF-g

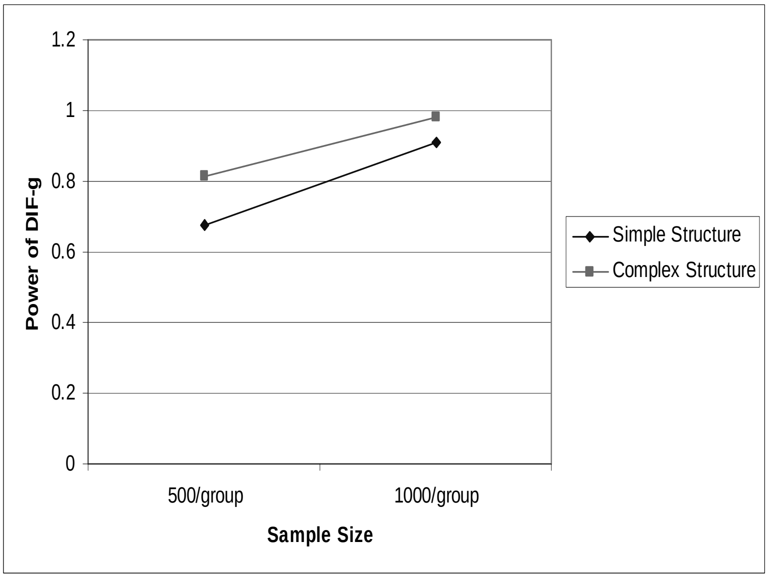


Figure 4.8: The Interaction between Ability Distribution and Q-matrix Structure on the Power of DIF-g

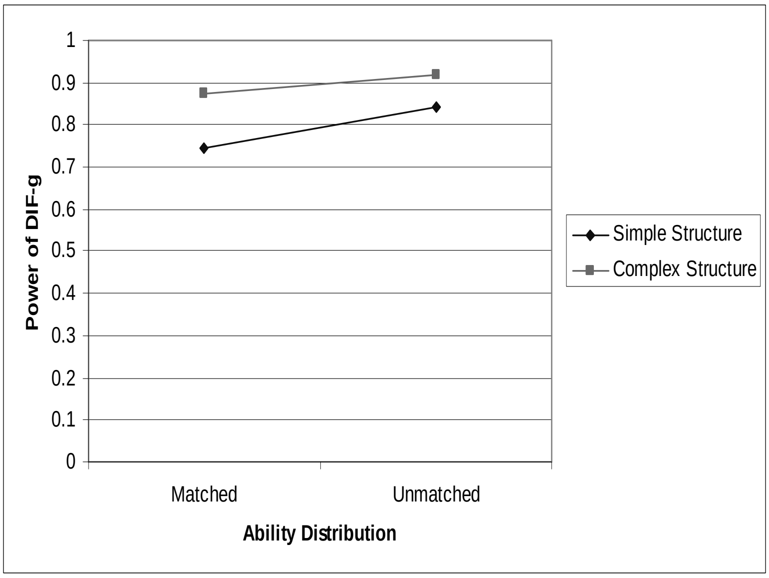
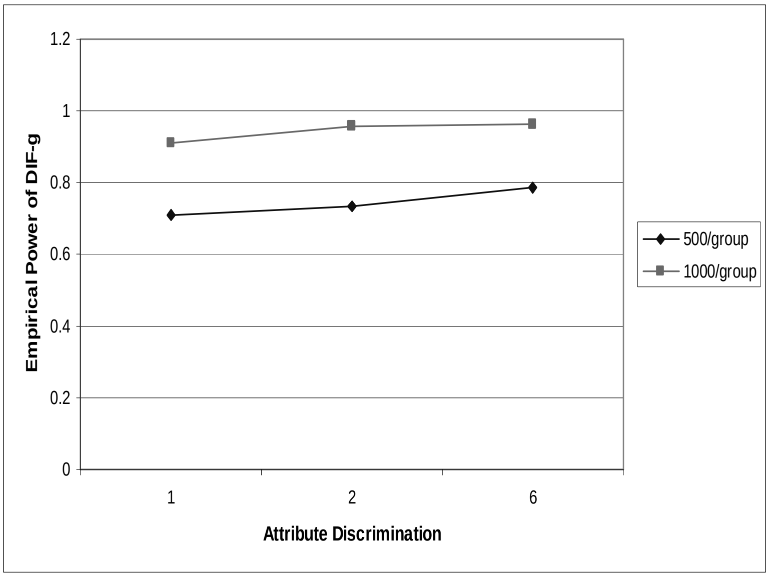


Figure 4.9: The Interaction between Attribute Discrimination and Sample Size on the Power of DIF-g



In Figure 4.10 and Figure 4.11, it can be seen that the difference in the power of DIF-s was smaller between simple structure and complex structure, when attribute discrimination increased to 6 or when ability distribution was matched. Figure 4.12 shows that the difference between uniform DIF and non-uniform DIF under large sample size was not as great as that under small sample size. Recall that DIF-s had higher power, with attribute discrimination $a = 6$, large sample size, matched ability distribution, or simple structure plus uniform DIF. These three figures showed that conditions under which power is good can reduce the impact of conditions under which power is poor.

Figure 4.10: The Interaction between Attribute Discrimination and Q-matrix Structure on the Power of DIF-s.

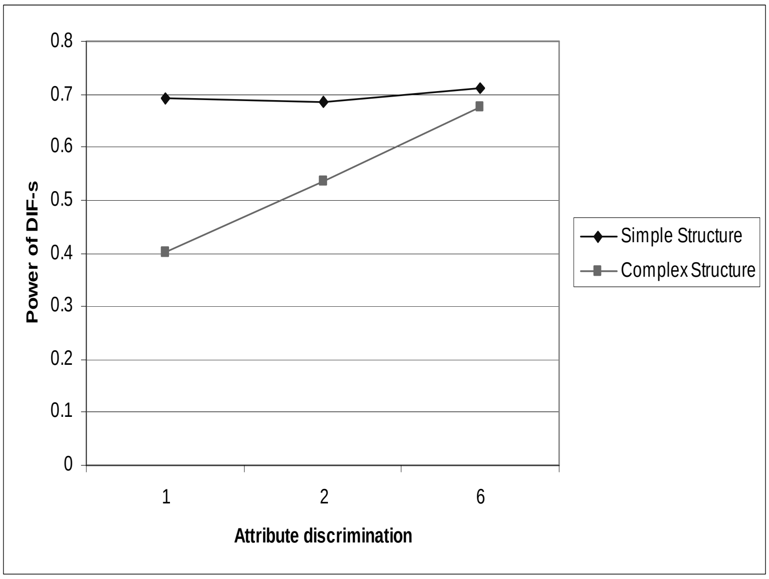


Figure 4.11: The Interaction between Ability Distribution and Q-matrix Structure on the Power of DIF-s

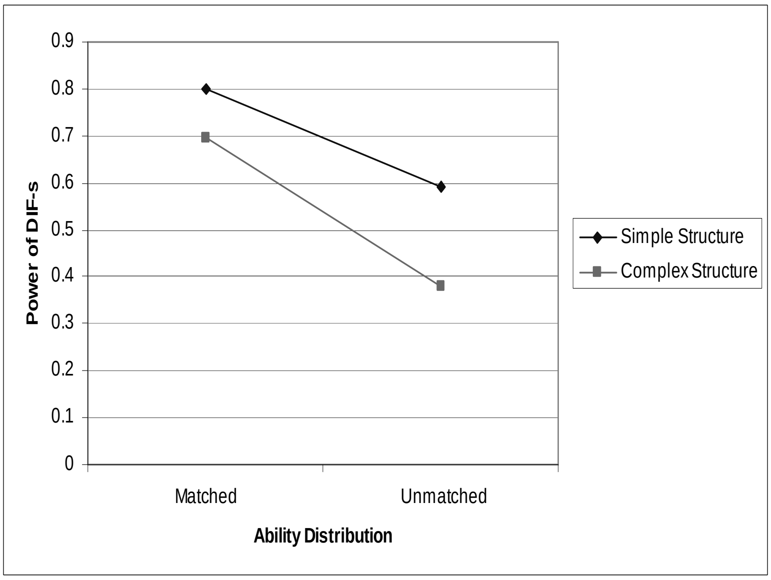
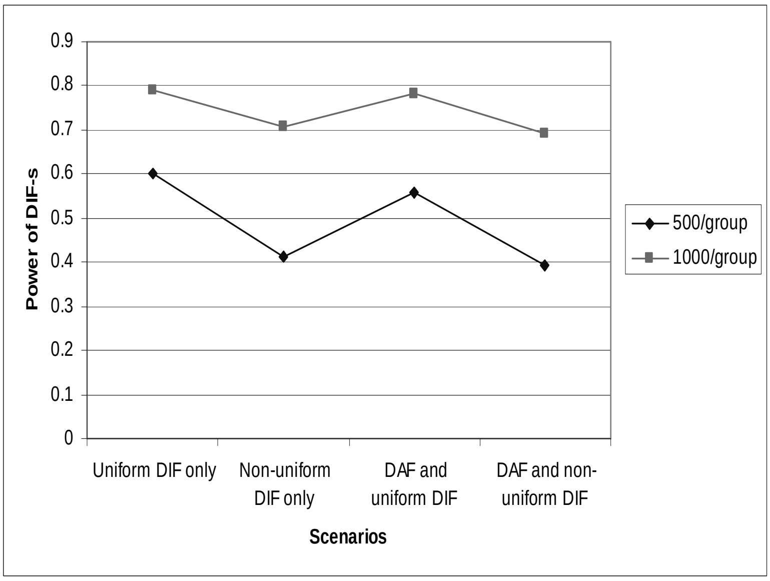


Figure 4.12: The Interaction between Scenarios and Sample Size on the Power of DIF-s



In contrast, Figure 4.13 and Figure 4.14 reflect a different pattern; that is, a condition under which power was good served to improve the power of a second condition under which power was also good. In Figure 4.13, for example, the power of DIF-s increased when simple structure was combined with uniform DIF over non-uniform DIF. Also, in Figure 4.14, the difference in power of DIF-s increased between matched and unmatched ability when higher attribute discrimination, i.e., either $a = 2$ or $a = 6$, was added.

Figure 4.13: The Interaction between Scenarios and Q-matrix Structure on the Power of DIF-s

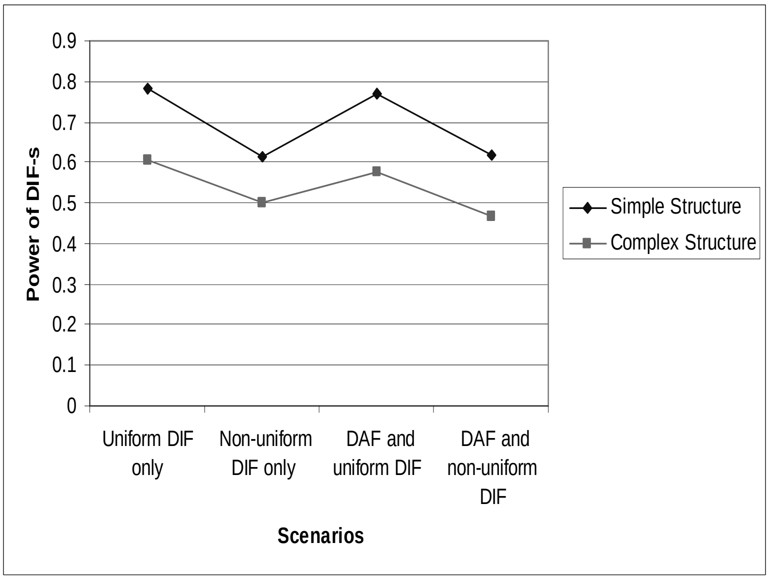
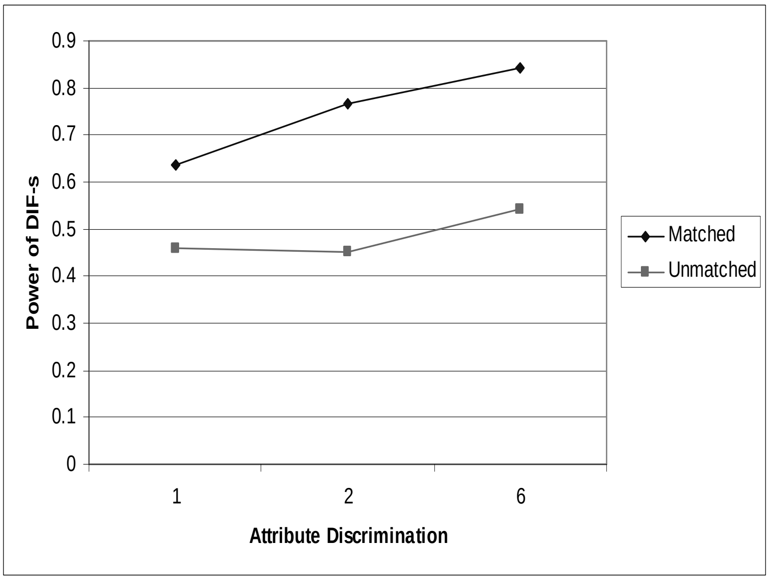


Figure 4.14: The Interaction between Attribute Discrimination and Ability Distribution on the Power of DIF-s



4.2.4 Comparison of DIF-g and DIF-s with MH DIF Detection

In this section, we compare detection of DIF with MH methods based on total scores and attribute mastery profiles as matching criteria. The MH method with total score as the matching criterion was done using the total raw score that included the studied item. These comparisons were done on the same simulated data and the Type I error and power were calculated under the same conditions as for the DIF-g and DIF-s results presented above. Since the MH method was not designed to detect DAF, these comparisons were only done for DIF detection. Tables 4.19 and 4.20 provided Type I error and power results for MH using the total score as the matching criteria. Table 4.21 and Table 4.22 provide Type I error and power for MH using the attribute mastery profiles as the matching criteria.

MH method with Totals Scores as a Matching Criterion. In Table 4.19, it can be seen that the Type I error rates were very large in those scenarios for which DAF was simulated. (These values are bolded in the table.) These error rates decreased, however, with the increase in attribute discrimination. This was consistent with expectation. Recall that two conditions need to be satisfied for DIF to be present: First, the item needs to be sensitive not only to the primary construct, but also to some secondary construct; second, a difference in the conditional distributions on the secondary construct needs to exist between groups of interest given a fixed value of primary construct (p. 2). When attribute discrimination was 1 or 2, the simulated data contained more multidimensionality, since lower attribute discrimination is equivalent to a lower correlation among attributes. As mentioned earlier in the description of the simulation study design, $a = 1$ or $a = 2$ made the data satisfy the first condition. In addition, DAF was defined as “a differential propensity of one group to have a greater probability of mastery on an attribute compared to another group, conditioned on general ability (p. 6).” Therefore, when DAF was generated in the data, the second condition was satisfied. Thus, those combinations of attribute discrimination of 1 or 2 and the conditions with DAF fully satisfied both conditions, and tended to produce DIF. This DIF was benign DIF since the secondary dimensions (i.e., those on which each of the attributes loaded) were auxiliary not nuisance dimensions. For this reason, the MH method with total scores as the matching criterion is likely to detect as DIF those items with benign DIF. As a result, the Type I error rates for these conditions are likely to be inflated. It can also be noticed that the bolded Type I error rates in Table 4.19 for complex structure are smaller than those for simple structure. This is because the response to one item could be determined by more than one attribute, the effect of one attribute with DAF on the studied item is smaller than that under simple structure where the response to one item is only determined by one attribute.

Since the Type I errors under the scenarios with DAF were large, we focus only on the power of MH for DIF detection in those scenarios without DAF. These are "Uniform DIF only" and "non-uniform DIF only" (see the bolded values in Table 4.20). The power for "uniform DIF only" were higher than for "non-uniform DIF only". The uniform DIF scenarios appear to have changed the probability of correct response in the same direction for all examinees in the focal group whereas the non-uniform DIF scenarios changed the probability of correct response in one direction for masters and the other direction for non-masters in the focal group. One possibility is that the DIF effect could have been cancelled under the non-uniform scenario, since the MH method is not able to detect non-uniform DIF. Although the power rates for the non-uniform scenario with complex structure, 1000 examinee per group, and unmatched ability distribution were around .7, this rate actually may be overestimated due to the inflated Type I error rate (see cells for corresponding conditions in Table 4.19). Likewise, one should use care in interpreting the relatively high power of MH for uniform DIF only, because the Type I errors of all conditions for uniform DIF were not controlled even based on Bradley's (1978) liberal criterion. As can been seen in Table 4.19, the Type I error rates for "Uniform DIF only" ranged from .080 to .232. These were not as high as for the scenarios with DAF, however. As a result, the power of uniform DIF might be somewhat overestimated. In general, though, one can probably conclude that the MH method with total score as the matching criterion had acceptable power for detecting uniform DIF, especially for the large sample size. Further, and as expected, the MH method failed to detect non-uniform DIF.

The model-based method developed in this study appeared to have reasonably good control of Type I errors for both DIF-g and DIF-s whether or not the scenarios simulated DAF. Moreover, the model-based method appeared to be able to detect non-uniform DIF by separating DIF into DIF-g for non-masters and DIF-s for masters. The power of DIF-g (see Table 4.16 and DIF-s (see Table 4.17) appears to be higher than the power of DIF (see Table 4.20 for the uniform DIF conditions. The Type I errors for DIF by the MH method were generally inflated compared to those for DIF-g and DIF-s even in the uniform DIF scenarios. In addition, when both Type I error of DIF-g and DIF-s were detected simultaneously, the total sample size for detecting DIF was reduced to the sample size for masters (for detecting DIF-s) and the sample size for non-masters (for detecting DIF-g). As a result, the power of either DIF-g or DIF-s was also reduced.

Table 4.19: Type I Error of DIF by MH Method Using Total Scores as a Matching Criterion

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	No DIF and DAF	.030	.054	.048	.040	.057	.044
		DAF only	**.400**	**.331**	**.121**	**.371**	**.179**	**.078**
		Uniform DIF only	.094	.104	.098	.092	.116	.088
		Non-uniform DIF only	.030	.050	.046	.046	.042	.050
		DAF and uniform DIF	**.398**	**.348**	**.160**	**.364**	**.240**	**.134**
		DAF and non-uniform DIF	**.406**	**.306**	**.112**	**.374**	**.200**	**.098**
	unmatched	No DIF and DAF	.030	.035	.048	.075	.081	.049
		DAF only	**.400**	**.297**	**.096**	**.222**	**.132**	**.062**
		Uniform DIF only	.080	.104	.102	.144	.150	.126
		Non-uniform DIF only	.042	.056	.062	.104	.116	.090
		DAF and uniform DIF	**.398**	**.322**	**.172**	**.258**	**.184**	**.138**
		DAF and non-uniform DIF	**.396**	**.282**	**.172**	**.254**	**.196**	**.118**
1000/g	matched	No DIF and DAF	.057	.070	.035	.054	.057	.043
		DAF only	**.412**	**.414**	**.217**	**.475**	**.433**	**.137**
		Uniform DIF only	.138	.162	.172	.194	.188	.158
		Non-uniform DIF only	.036	.072	.038	.078	.070	.050
		DAF and uniform DIF	**.446**	**.476**	**.282**	**.460**	**.434**	**.262**
		DAF and non-uniform DIF	**.426**	**.444**	**.192**	**.428**	**.446**	**.180**
	unmatched	No DIF and DAF	.044	.040	.043	.089	.067	.060
		DAF only	**.433**	**.411**	**.136**	**.272**	**.262**	**.131**
		Uniform DIF only	.120	.126	.172	.220	.232	.214
		Non-uniform DIF only	.034	.046	.070	.150	.164	.128
		DAF and uniform DIF	**.514**	**.450**	**.292**	**.294**	**.280**	**.238**
		DAF and non-uniform DIF	**.450**	**.400**	**.158**	**.290**	**.280**	**.192**

Table 4.20: Power of DIF by MH Method Using Total Scores as a Matching Criterion

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	Uniform DIF only	**.79**	**.84**	**.89**	**.82**	**.86**	**.86**
		Non-uniform DIF only	**.02**	**.04**	**.04**	**.26**	**.16**	**.06**
		DAF and uniform DIF	.73	.65	.75	.72	.74	.84
		DAF and non-uniform DIF	.36	.18	.05	.51	.31	.10
	unmatched	Uniform DIF only	**.74**	**.73**	**.78**	**.77**	**.86**	**.80**
		Non-uniform DIF only	**.09**	**.18**	**.22**	**.50**	**.45**	**.36**
		DAF and uniform DIF	.62	.57	.69	.71	.73	.74
		DAF and non-uniform DIF	.34	.26	.24	.66	.58	.38
1000/g	matched	Uniform DIF only	**.97**	**.97**	**.99**	**1.00**	**1.00**	**1.00**
		Non-uniform DIF only	**.03**	**.06**	**.02**	**.58**	**.38**	**.13**
		DAF and uniform DIF	.89	.80	.89	.80	.80	.92
		DAF and non-uniform DIF	.40	.33	.09	.84	.67	.21
	unmatched	Uniform DIF only	**.92**	**1.00**	**1.00**	**.98**	**.98**	**.97**
		Non-uniform DIF only	**.11**	**.33**	**.40**	**.73**	**.74**	**.62**
		DAF and uniform DIF	.81	.80	.96	.82	.83	.91
		DAF and non-uniform DIF	.42	.40	.44	.81	.74	.60

MH Method with Attribute Profiles as a Matching Criterion. In Table 4.21, results were reported for the MH method using attribute profiles as the matching criterion. The error rates in Table 4.21 were not abnormally high for the scenarios with DAF as was observed (above) for the MH method using total scores as the matching criterion. The conditions with attribute discrimination $a = 1$ or $a = 2$, however, did consistently have inflated Type I error rates than when attribute discrimination was simulated as $a = 6$. This result was consistent with the results reported by Zhang (2007). Zhang found Type I error using the MH method and attribute profiles as the matching criterion to have decreased with an increase in the correlation among attributes. The increase in correlation among attributes is equivalent to an increase in attribute discrimination. As the attribute profiles were used as the matching criterion, the result was that the MH was sensitive to the correctness of classification of mastery status. As indicated in Section 4.2.1, the correct classification rate of attribute mastery increased with an increase in attribute discrimination. This could be a possible reason for the reduced Type I error rates in conditions with higher attribute discrimination.

In Table 4.22, the power was highlighted for the conditions with attribute discrimination $a = 6$, since Type I error rates were relatively controlled under these same conditions. The power showed the same pattern as the power based on MH method using total scores as the matching criterion: the power rates were high for uniform DIF under all conditions, but much lower for non-uniform DIF. That is, both MH methods were not capable of detecting non-uniform DIF.

The MH method using attribute profiles as the matching criterion appeared to have avoided the extent of loss of Type I error control that appeared in those conditions in which DAF was simulated. Even so, the Type I error control was not as good as the model-based method for conditions with attribute discrimination $a = 1$ or $a = 2$. The MH method using attribute profiles as matching criterion, however, was more powerful at detecting DIF, although detection of non-uniform DIF was not good.

Table 4.21: Type I Error of DIF by MH Method Using Attribute Profiles as a Matching Criterion

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	No DIF and DAF	.078	.108	.083	.120	.083	.060
		DAF only	.128	.100	.080	.188	.115	.059
		Uniform DIF only	.082	.094	.080	.124	.084	.060
		Non-uniform DIF only	.082	.124	.086	.128	.090	.064
		DAF and uniform DIF	.076	.100	.080	.186	.122	.072
		DAF and non-uniform DIF	.088	.176	.086	.218	.112	.064
	unmatched	No DIF and DAF	.115	.115	.075	.188	.088	.067
		DAF only	.128	.120	.078	.187	.104	.054
		Uniform DIF only	.134	.114	.078	.174	.078	.074
		Non-uniform DIF only	.160	.136	.086	.200	.102	.056
		DAF and uniform DIF	.132	.118	.080	.186	.084	.052
		DAF and non-uniform DIF	.150	.146	.092	.218	.094	.062
1000/g	matched	No DIF and DAF	.094	.102	.046	.097	.107	.054
		DAF only	.110	.128	.048	.161	.153	.060
		Uniform DIF only	.086	.118	.058	.106	.102	.060
		Non-uniform DIF only	.102	.138	.070	.128	.114	.064
		DAF and uniform DIF	.114	.120	.064	.156	.142	.064
		DAF and non-uniform DIF	.154	.138	.064	.216	.186	.074
	unmatched	No DIF and DAF	.126	.116	.059	.214	.105	.059
		DAF only	.147	.097	.054	.217	.097	.056
		Uniform DIF only	.128	.098	.056	.212	.108	.050
		Non-uniform DIF only	.168	.124	.058	.248	.110	.064
		DAF and uniform DIF	.164	.088	.054	.214	.108	.052
		DAF and non-uniform DIF	.156	.120	.054	.324	.124	.050

Table 4.22: Power of DIF by MH method Using Attribute Profiles as a Matching Criterion

			Simple Structure			Complex Structure		
			a=1	a=2	a=6	a=1	a=2	a=6
500/g	matched	Uniform DIF only	.96	.99	**.97**	.92	.96	**.96**
		Non-uniform DIF only	.12	.09	**.05**	.42	.22	**.04**
		DAF and uniform DIF	.97	.93	**.96**	.75	.89	**.97**
		DAF and non-uniform DIF	.11	.20	**.04**	.28	.15	**.03**
	unmatched	Uniform DIF only	.92	.96	**.92**	.96	.80	**.93**
		Non-uniform DIF only	.21	.35	**.35**	.71	.61	**.44**
		DAF and uniform DIF	.89	.93	**.93**	.91	.90	**.92**
		DAF and non-uniform DIF	.23	.34	**.36**	.74	.68	**.46**
1000/g	matched	Uniform DIF only	1.00	1.00	**1.00**	1.00	1.00	**1.00**
		Non-uniform DIF only	.08	.15	**.05**	.60	.36	**.12**
		DAF and uniform DIF	.99	.99	**1.00**	.86	.96	**1.00**
		DAF and non-uniform DIF	.22	.22	**.09**	.42	.28	**.11**
	unmatched	Uniform DIF only	1.00	1.00	**1.00**	1.00	1.00	**1.00**
		Non-uniform DIF only	.23	.47	**.62**	.80	.88	**.77**
		DAF and uniform DIF	1.00	1.00	**1.00**	.99	.98	**.99**
		DAF and non-uniform DIF	.42	.53	**.61**	.74	.80	**.76**

4.3 Detection of DAF and DIF on a Statewide Mathematics Test

To illustrate the use of the model-based method for DAF and DIF detection, the modified higher-order DINA model was applied to a gender DIF problem using data sampled from a statewide mathematics test. Gender DAF detection focuses on identifying weakness and strength for males and females given the same ability level, whereas gender DIF detection seeks to identify group differences conditioning on attribute mastery. In addition, model-based DIF detection was also compared with two MH methods, one with total scores and a second with attribute profiles as matching criteria.

4.3.1 Data Description

A sample of 2,000 examinees (993 males and 1007 females) was randomly drawn from a total statewide sample of 136,156 students in Grade 3 who took the 2003 Florida Comprehensive Assessment Test (FCAT) Mathematics Test (Florida Department of Education, 2003). Before drawing the sample, the examinees were excluded if they had received any accommodation, had an indication of any primary exceptionality, or were identified as limited English Proficient.

The test included 40 operational multiple-choice items, designed to measure one of the following five content strands: Number Sense and Operation, Measurement, Geometry and Spatial Sense, Algebraic Thinking, Data Analysis and Probability. Six non-operational item locations were present on each of the 10 forms of the test, and were used to field test new or revised items. Items in these locations were not analyzed in this study. In this study, the content strands were treated as attributes. A 40 $\times$ 5 Q-matrix was created from the item and task specification for the FCAT (see Table 4.23). The resulting Q-matrix had a simple structure as each item was designed to measure only a single attribute.

Table 4.24 presents the descriptive statistics for the whole test and for each attribute for females and for males. The data analyses were done using the same software as used for the simulation study. It can be seen the mean total score for males was about 1 unit higher than that for females. For each attribute, the mean score on each attribute was also a little higher for males than females. Results with an independent samples t-test, however, suggested all differences except for Geometry were significant difference due to the large sample size of the data ($N = 2000$).

Table 4.23: Q-Matrix for FCAT 2003 Grade 3 Mathematics Test

Items	Number Sense & Operation	Measurement	Geometry & Spatial Sense	Algebraic Thinking	Data Analysis &Probability
1	1	0	0	0	0
2	0	0	0	1	0
3	0	0	0	0	1
4	0	1	0	0	0
5	0	0	0	0	1
6	1	0	0	0	0
7	1	0	0	0	0
8	0	1	0	0	0
9	0	0	1	0	0
10	0	1	0	0	0
11	0	0	1	0	0
12	0	0	1	0	0
13	1	0	0	0	0
14	0	0	0	1	0
15	1	0	0	0	0
16	0	0	0	0	1
17	0	0	0	1	0
18	1	0	0	0	0
19	0	0	0	0	1
20	0	0	1	0	0
21	0	0	0	0	1
22	0	0	1	0	0
23	0	0	1	0	0
24	0	1	0	0	0
25	0	1	0	0	0
26	0	0	0	1	0
27	0	0	1	0	0
28	1	0	0	0	0
29	1	0	0	0	0
30	0	0	0	1	0
31	1	0	0	0	0
32	1	0	0	0	0
33	1	0	0	0	0
34	1	0	0	0	0
35	0	0	0	1	0
36	0	0	0	0	1
37	0	0	0	0	1
38	0	1	0	0	0
39	0	1	0	0	0
40	0	1	0	0	0

Table 4.24: Descriptive Sample Statistics for FCAT 2003 Grade 3 Mathematics

		Total Score	Number Operation	Measurement	Geometry	Algebraic Thinking	Data Analysis
Female	Mean	23.41	6.62	5.35	4.21	3.18	4.05
	(S.D.)	(7.25)	(2.77)	(1.72)	(1.59)	(1.45)	(1.67)
Male	Mean	24.48	6.99	5.66	4.24	3.39	4.18
	(S.D.)	(7.48)	(2.78)	(1.71)	(1.66)	(1.50)	(1.69)
t test	t	-3.25	-3.04	-4.12	-.46	-3.24	-1.72
	(p-value)	(.001)	(.002)	(.000)	(.646)	(.001)	(.085)

4.3.2 Results for Real Data Example

The male group was arbitrarily chosen as the reference group. Item parameter estimates for both female and male group are presented in Table 4.25. The slip parameters were generally low, and some of the guessing parameters appeared high and, as a result, possibly problematic: Specifically, 15 of 40 guessing parameters were higher than .50. Typically, high guessing parameter estimates occur when more attributes have been specified for an item than necessary, although for this test, only one attribute was specified for each item. Another possible reason for the high guessing parameters estimated in the data may be that the attributes were inaccurately specified for those items or the difficulties of those items were much lower than the ability levels of students (i.e., were too easy). As a result, more students were classified as masters. Since high guessing occurred in both groups, it may not affect DAF and DIF detection so much.

Table 4.26 provides the frequencies and percentages of each attribute mastery pattern for total group, females and males. In principle, five attributes will generate $2^5 = 32$ attribute mastery patterns. However, in this real data example, only 27 of these patterns occurred. Of those patterns that were present, the patterns “00000” and “11111” accounted for almost 80% of the patterns in the total sample. The third largest pattern was “01000”, and the percentages of other patterns were all lower than 2%. Similar distributions were found for female and male groups. These kinds of distribution indicated the five attributes were highly correlated. That is, if a student failed to master one attribute, the same student tended to not master all attributes. The results also suggest that the second attribute might be the easiest one since 6.2% of examinees mastered this one but failed all the others.

Table 4.25: Item Parameter Estimates for both Females and Males

Items	g_{female}(S.E.)	g_{male}(S.E.)	s_{female}(S.E.)	s_{male}(S.E.)
1	0.90(0.01)	0.86(0.01)	0.01(0.01)	0.00(0.00)
2	0.72(0.02)	0.63(0.02)	0.03(0.01)	0.04(0.01)
3	0.68(0.02)	0.69(0.02)	0.22(0.02)	0.22(0.01)
4	0.55(0.02)	0.57(0.02)	0.25(0.02)	0.30(0.01)
5	0.43(0.02)	0.38(0.02)	0.08(0.01)	0.11(0.01)
6	0.46(0.02)	0.43(0.02)	0.07(0.01)	0.11(0.01)
7	0.32(0.02)	0.32(0.02)	0.19(0.02)	0.23(0.02)
8	0.38(0.02)	0.26(0.02)	0.22(0.02)	0.31(0.02)
9	0.40(0.02)	0.40(0.02)	0.21(0.02)	0.22(0.02)
10	0.81(0.02)	0.70(0.02)	0.06(0.01)	0.08(0.01)
11	0.37(0.02)	0.34(0.02)	0.35(0.02)	0.43(0.02)
12	0.54(0.02)	0.54(0.02)	0.15(0.02)	0.14(0.01)
13	0.57(0.02)	0.54(0.02)	0.12(0.01)	0.16(0.01)
14	0.25(0.02)	0.23(0.02)	0.33(0.02)	0.36(0.02)
15	0.40(0.02)	0.35(0.02)	0.20(0.01)	0.19(0.01)
16	0.23(0.02)	0.10(0.01)	0.34(0.02)	0.52(0.02)
17	0.45(0.02)	0.52(0.02)	0.27(0.02)	0.21(0.02)
18	0.24(0.02)	0.22(0.02)	0.26(0.02)	0.34(0.02)
19	0.51(0.02)	0.44(0.02)	0.09(0.01)	0.11(0.01)
20	0.31(0.02)	0.38(0.02)	0.35(0.02)	0.37(0.02)
21	0.24(0.02)	0.25(0.02)	0.47(0.02)	0.47(0.02)
22	0.59(0.02)	0.58(0.02)	0.14(0.01)	0.14(0.01)
23	0.51(0.02)	0.57(0.02)	0.16(0.02)	0.18(0.01)
24	0.58(0.02)	0.55(0.03)	0.05(0.01)	0.05(0.00)
25	0.57(0.02)	0.57(0.02)	0.10(0.01)	0.10(0.01)
26	0.41(0.02)	0.42(0.02)	0.27(0.02)	0.24(0.02)
27	0.42(0.02)	0.40(0.02)	0.29(0.02)	0.31(0.02)
28	0.44(0.02)	0.43(0.02)	0.24(0.02)	0.28(0.02)
29	0.41(0.02)	0.36(0.02)	0.07(0.01)	0.10(0.01)
30	0.13(0.01)	0.09(0.01)	0.34(0.02)	0.52(0.02)
31	0.20(0.01)	0.17(0.01)	0.53(0.02)	0.59(0.02)
32	0.21(0.02)	0.19(0.01)	0.28(0.02)	0.29(0.02)
33	0.28(0.02)	0.22(0.01)	0.38(0.02)	0.38(0.02)
34	0.20(0.02)	0.24(0.01)	0.50(0.02)	0.44(0.02)
35	0.39(0.02)	0.36(0.02)	0.32(0.02)	0.37(0.02)
36	0.20(0.02)	0.26(0.02)	0.26(0.02)	0.23(0.02)
37	0.55(0.02)	0.59(0.02)	0.10(0.01)	0.12(0.01)
38	0.18(0.02)	0.09(0.02)	0.31(0.02)	0.43(0.02)
39	0.62(0.02)	0.49(0.02)	0.13(0.01)	0.21(0.01)
40	0.45(0.02)	0.49(0.02)	0.16(0.01)	0.20(0.01)

Table 4.26: Frequencies and Percentages of Attribute Mastery Patterns

	Total		Female		Male	
Profiles	Frequency	(Percent)	Frequency	(Percent)	Frequency	(Percent)
00000	719	(36.0)	351	(34.9)	368	(37.1)
00001	8	(.4)	4	(.4)	4	(.4)
00011	2	(.1)	0	(0.0)	2	(.2)
00100	6	(.3)	3	(.3)	3	(.3)
00101	1	(.0)	0	(0.0)	1	(.1)
01000	125	(6.2)	79	(7.8)	46	(4.6)
01001	33	(1.6)	23	(2.3)	10	(1.0)
01010	11	(.6)	3	(.3)	8	(.8)
01011	6	(.3)	3	(.3)	3	(.3)
01100	12	(.6)	8	(.8)	4	(.4)
01101	22	(1.1)	16	(1.6)	6	(.6)
01110	3	(.2)	0	(0.0)	3	(.3)
01111	18	(.9)	10	(1.0)	8	(.8)
10000	10	(.5)	2	(.2)	8	(.8)
10001	3	(.2)	1	(.1)	2	(.2)
10010	1	(.0)	0	(0.0)	1	(.1)
10101	2	(.1)	0	(0.0)	2	(.2)
10110	2	(.1)	0	(0.0)	2	(.2)
10111	3	(.2)	0	(0.0)	3	(.3)
11000	19	(1.0)	7	(.7)	12	(1.2)
11001	18	(.9)	11	(1.1)	7	(.7)
11010	12	(.6)	4	(.4)	8	(.8)
11011	33	(1.6)	11	(1.1)	22	(2.2)
11100	7	(.4)	5	(.5)	2	(.2)
11101	31	(1.6)	21	(2.1)	10	(1.0)
11110	18	(.9)	7	(.7)	11	(1.1)
11111	874	(43.7)	438	(43.5)	436	(43.9)
Total	2000	(100.0)	1007	(100.0)	993	(100.0)

4.3.3 DAF Detection

Table 4.27 presented all attribute parameter estimates and their corresponding 95% credibility interval for the higher level model. This included the DAF index parameters γ^{adj}. In Table 4.27, estimates of the attribute discrimination parameter can be seen to be as high as 6. The β^{adj} estimates represented the attribute difficulties for the reference group. The second attribute, Measurement, was the easiest and the third attribute, Algebraic Thinking, was the hardest. These results explain the distribution of attribute patterns observed in Table 4.26. High attribute discrimination parameter estimates suggested each attribute was highly correlated to general ability. As a result, the all-non-mastery pattern "00000" and the all-mastery pattern "11111" were the dominant patterns. In addition, since the second attribute was the easiest, the pattern "01000" had a greater probability of occurring than other patterns containing the second attribute being non-mastered. The 95% credibility interval on γ^{adj} indicated the second attribute, Measurement, favored the focal group (female group), the fourth attribute, Algebraic Thinking, favored the reference group (male group).

Table 4.27: Attribute Parameter Estimates and 95% CIs

Parameters	Means	95% CIs
a	6.80	(6.04, 7.54)
$\beta 1^{adj}$	.03	(-0.45, 0.49)
$\beta 2^{adj}$	-1.21	(-1.85, -0.59)
$\beta 3^{adj}$	.66	(-0.13, 1.37)
$\beta 4^{adj}$	.41	(-0.22, 1.04)
$\beta 5^{adj}$	.16	(-0.35, 0.69)
$\gamma 1^{adj}$	.40	(-0.27, 1.08)
$\gamma 2^{adj}$	-.89	(-1.70, -0.09)
$\gamma 3^{adj}$	-.14	(-1.12, 0.79)
$\gamma 4^{adj}$	.86	(0.02, 1.77)
$\gamma 5^{adj}$	-.23	(-0.92, -0.23)

In Table 4.28, the group ability means and group marginal attribute mastery proportions are presented. It can be seen that ability was similar between gender groups as females were lower than males by only .03. Females and males differed, however, in the mastery of different attributes: Females had a higher mastery proportion for Measurement and males had a higher mastery proportion for Algebraic Thinking. Both differences were about 3.8%, and group differences for the other attributes were small.

Table 4.28: General Ability Mean and Marginal Attribute Mastery Proportions

	General Ability	Number Operation	Measurement	Geometry	Algebraic Thinking	Data Analysis
Male	0	52.97%	60.02%	49.45%	51.06%	51.96%
Female	-.03	50.35%	64.15%	50.45%	47.27%	53.43%

4.3.4 DIF Detection

Results were given in Table 4.29 for the MH with total score and with attribute profile as matching criteria. Recall that a significant χ^2 for the MH method indicates uniform DIF. In addition, estimates of Δg and Δs for the model-based method as well as the 95% credibility interval for DIF-g and DIF-s detection were given in Table 4.29. If the 95% CI on Δg and Δs does not contain 0, this suggests a significant DIF-g or DIF-s. A difference in the signs for DIF-g and DIF-s suggest uniform DIF. Similarly, the same sign for DIF-g and DIF-s suggests non-uniform DIF. This is because an increase in both g and s will result in an increase in the proportion of correct responses for non-masters, and a decrease in that for masters, respectively. Conversely, a decrease in both g and s will result in a decrease in the proportion of correct responses for non-masters but an increase in the proportion of masters, respectively. Finally, an increase in g and a decrease in s or a decrease in g and an increase in s will result in an increase or decrease in the proportion of correct responses for both masters and non-masters, respectively.

It can be seen in the table that MH with the total score as matching criterion was able to detect 15 uniform DIF items, and MH with the attribute profile as the matching criterion detected 17 uniform DIF items. There were 11 items with one or both DIF-g or DIF-s detected. All of these indicated uniform DIF as the signs on Δg and Δs were different.

This result was consistent with results from the comparisons among the three methods in the simulation study. Recall that those results indicated the model-based method was more conservative and yielded lower Type I error rates for detection of DIF, even though eight common items were detected as DIF items by all three methods. In this data set, however, DIF detection results appeared to be similar among the three methods. The similarity in results was possibly due to the high correlations among attributes. As noted earlier, Type I errors are reduced for both the MH with attribute profile and MH with total score matching criteria, when attributes are highly correlated. The model-based method was more powerful than the other two MH methods, when non-uniform DIF existed, although they all performed similarly, when the data were unidimensional and no non-uniform DIF existed. One possible reason the MH with the attribute profile detected a relatively large number of DIF items is that sparseness occurred in many attribute mastery patterns.

Table 4.29: DIF Detection based on the Three Methods

		MH with		MH with		Model Based Method			
		Total Score		Attribute Profile		DIF-g		DIF-s	
Item	Skill	χ^2(p value)	$\Delta\alpha_{MH}$	χ^2(p value)	$\Delta\alpha_{MH}$	Δg(95%CI)	$\Delta\alpha_{MH}$	Δs(95%CI)	$\Delta\alpha_{MH}$
2	4	5.52(.019)	-0.74	9.10(.003)	-0.93	-.09(-.15,-.02)	-0.96		
5	5			10.00(.002)	-0.88				
6	1			3.869(.049)	-0.54				
8	2	6.39(.011)	-0.61	23.53(.000)	-1.18	-.12(-.19,-.05)	-1.30	.09(.03,.15)	-1.09
10	2	6.04(.014)	-0.81	13.80(.000)	-0.32	-.11(-.17, -.04)	-1.41	.08(.01,.15)	
11	3			5.49(.019)	-0.52				-0.81
13	1			8.77(.003)	-0.76				
16	5	40.24(.000)	**-1.58**	61.49(.000)	**-1.95**	-.13(-.18,-.08)	**-2.22**	.18(.11,.25)	**-1.74**
17	4	13.35(.000)	0.86	7.63(.006)	0.65	.08(.01,.14)	0.71		
18	1			5.73(.017)	-0.59			.08(.01,.15)	-0.85
19	5			4.39(.036)	-0.57	-.07(-.14,-.002)	-0.69		
20	3	4.37(.037)	0.48						
21	5	5.17(.023)	0.54						
23	3	7.90(.005)	0.69						
25	2	4.06(.044)	0.58						
29	1			6.09(.014)	-0.7				
30	4	22.91(.000)	**-1.58**	34.23(.000)	**-1.54**			.18(.11,.26)	**-1.75**
31	1			4.82(.028)	-0.53				
34	1	11.83(.001)	0.83	4.899(.027)	0.53				
36	5								
37	5	5.34(.021)	0.62						
38	2	8.89(.003)	-0.74	41.34(.000)	**-1.65**	-.08(-.15, -.02)	**-1.64**	.13(.07,.19)	-1.28
39	2	12.34(.000)	-0.90	21.28(.000)	-1.17	-.13(-.20,-.05)	-1.20	.08(.03,.13)	-1.31

Although a lot of items were detected as DIF items, many of them were negligible or moderate according to the ETS standards used in this study, when $|\Delta\alpha_{MH}| \geq 1.5$). Results for both MH methods were in agreement that only three items (Items 16, 30, and 38) had large DIF. For DIF-g and DIF-s, no specific criterion was developed for this study, but some calculations can be approximated to obtain a useful result for $\Delta\alpha_{MH}$ using equation 2.9: In that equation, R_{rk} and W_{rk} are the counts of right and wrong responses in the reference group at score level k, R_{fk} and W_{fk} are the counts of right and wrong responses in the focal group at level k, and N_{tk} is the number of examinees in the total group at level k $(k = 1, \ldots, K)$. In the CDM in this study, only two levels existed for each item, masters or non-masters of that attribute. For DIF-g, since only non-masters are considered, the following is used:

$$\alpha_{MH} = \frac{g_r(1-g_f)}{(1-g_r)g_f} \tag{4.1}$$

where g_r, g_f are the guessing parameters for the reference group and focal groups, respectively. For DIF-s, since only masters are considered, the following equation is used:

$$\alpha_{MH} = \frac{s_f(1-s_r)}{(1-s_f)s_r} \tag{4.2}$$

where s_r and s_f are the slip parameters for the reference group and focal groups, respectively. After this transformation, the same criteria for $\Delta\alpha_{MH}$ could be used for DIF-g and DIF-s. Using this approach, all three methods appeared to have detected the same three items with large DIF. These three items measured attributes "Data analysis", "Algebraic thinking" and "Measurement" respectively. Currently, few if any tests have been written to conform to a Q-matrix. As a result, it is possible that the Q-matrix that was developed in this example, given the existing items, may have inadvertently resulted in more highly correlated attributes than would might have occurred were a test to be constructed to specifically measure attributes as in a predetermined Q-matrix. This reality was essentially a limitation when trying to apply a DIF detection method to real data in the context of a cognitive diagnostic modeling framework. Here both the model-based method in this study and the MH with attribute profiles used by Zhang (2007) were developed specifically for cognitive diagnostic modeling assessment, but both failed to show the advantage of the methods, when applied to real data from a test constructed to other than cognitive diagnostic modeling framework specifications.

Chapter 5

Discussion

This dissertation presented a modified higher-order DINA model for separating the source of construct-relevant (i.e. benign) DIF from construct irrelevant (i.e., adverse) DIF. This model-based method provides a natural framework for detecting both differential attribute functioning (DAF) and differential item functioning (DIF) in a cognitive diagnostic modeling framework: The higher level IRT model provides an estimate of group difference in attribute difficulty as an index of DAF and the lower level DINA model provides an estimate of group difference in item parameters with an index of DIF incorporated into the model. DIF detection ensures test fairness and improves test validity in terms of group difference in item performance after conditioning attributes mastery profiles, whereas DAF detection provides a good understanding of group strength and weakness in terms of a set of cognitive attributes after conditioning on general ability.

An MCMC algorithm employing Gibbs sampling was used to estimate the new model, and a simulation study was done to examine model recovery, Type I error rates, and power under practical testing conditions. There were five factors manipulated in the simulation study: Q-matrix structure, attribute discrimination parameters, sample size, ability distribution difference, scenarios of DIF and DAF combination. For DIF detection, the model-based method was also compared with the MH method using two types of matching criteria, a total score as the matching criterion and an attribute profile as the matching criterion. Finally, a statewide mathematics test was used to illustrate the implementation and possible limitations of the new method.

5.0.5 Summary of Simulation Study Results

The recovery of item parameters was generally better than the recovery of attribute parameters. One reason this occurred may be that the attribute level was not as informative as the item level due to the limited number of attributes. Specifically, the attribute discrimination estimates were biased when the generating value was 6. One possible explanation for the poorer recovery may be because of the way Bayesian estimation obtains the posterior distribution. It does this by combining the prior distribution with the likelihood. When less information is provided by the data (i.e., the likelihood), the posterior distribution weighs more heavily in the prior. The result is that shrinkage toward the mean of the prior occurs in the estimates of attribute discrimination. In terms of the problem in this study, more robust estimates of general ability recovery may have resulted had more attributes been simulated and had more items been simulated as measuring each of the attributes. However, in reality, few tests measure more than 10 different attributes. Therefore, there is a gap between what would be a better model from the standpoint of the model developed in this study and what is generally available in the usual testing program that would improve the estimation of attribute parameters and on DAF detection.

Type I error and power were calculated to assess the effects of different testing conditions on both DAF and DIF detection by manipulating the following factors: Q-matrix structure, attribute discrimination parameters, sample size, ability distribution difference, scenarios of DIF and DAF combination. Type I error for DIF tests was evaluated using the liberal range suggested by Bradley (1978). This was not the case for the DAF tests, however, as many of the conditions yielded Type I error for DAF that were out of that range. Most of these were deflated. The combination of small numbers of replications and small numbers of attributes appears to be a possible reason for the relatively high degree of variability of Type I error control for DAF. In particular, more replications would seemed to be required to obtain more stable estimates of Type I error in DAF. In addition, Type I error of both DAF and DIF seemed not to be very sensitive to the five manipulated factors. Only Type I error of DIF-s was sensitive to sample size and attribute discrimination.

Unlike Type I error, the power of both DAF and DIF (DIF-g and DIF-s) varied across different testing conditions. The pattern of variability was consistent with the pattern of recovery of DAF- and DIF-related parameters: The conditions with higher power for DAF responded to the conditions with lower RMSE of γ (the DAF parameter), and the conditions with higher power in DIF-g and DIF-s responded to the conditions with lower RMSE of g and s.

Clearly, the power of both DAF and DIF was higher in the large sample. Even though the power rates of DIF-g and DIF-s were more dependent on the sample size of non-masters and masters, respectively, they likewise increased markedly. This is because the number of masters and non-masters increased with the increase in total sample size. The four other factors did not function consistently in terms of the power between DAF and DIF, or between DIF-g and DIF-s.

High attribute discrimination ($a = 6$) resulted in more bias in the estimation of attribute parameters. Thus, the power rate of DAF dropped, when attribute discrimination increased to 6. The power was similar, however, when discrimination was simulated to be $a = 1$ and $a = 2$. High attribute discrimination, however, did result in improved power of DIF-g and DIF-s, most likely because attributes with higher discrimination can distinguish master and non-master more easily. As a result, more examinees were correctly classified as masters or non-masters. Further, high attribute discrimination appeared to help both the recovery of s and g, as well as the power of DIF-g and DIF-s. To explain the effect of high attribute discrimination, it is useful to first note that highly correlated attributes improved the correct classification of masters and non-masters. When attributes are highly correlated, the number of profile patterns that are present in the data tend to be reduced. In the extreme, that is, with perfectly correlated attributes, it is possible that the number of patterns might actually reduce to two, all 0's (non-masters) and all 1's (masters). Although this might make classification somewhat easier, the assessment would likely turn out to be less informative.

The different scenarios constructed in the simulation study were mainly formed with different combinations of DAF, DIF-s and DIF-g. The purpose of generating these different scenarios was to help determine under which conditions the model developed in this study could detect DAF and DIF simultaneously. Results suggested that the performance of DAF, DIF-g and DIF-s detection based on the model was not greatly affected by most of the scenarios. The exception was that the power rate of DIF-s was lower for non-uniform DIF than uniform DIF. This likely occurred because non-uniform DIF had higher slip parameters than uniform DIF in the simulation design and higher slip results in less precise estimation of masters.

In those conditions with simple structure and matched ability distributions between reference and focal groups, the power of DAF and DIF-s increased as expected, but the power of DIF-g decreased. The manipulation of simple structure or complex structure, and matched or unmatched ability distribution did appear to influence the number of masters and non-masters given sample size. As explained in Chapter 4, the condition of simple structure and matched ability distribution generated approximately equal numbers of masters and non-masters, and the complex structure and unmatched ability distribution generated more non-masters than masters. Thus, in the complex Q-Matrix structure and unmatched ability condition, the detection of DIF-g was better.

The comparison of the model-based method developed in this dissertation with the other two MH methods for DIF detection showed the model-based method had better Type I error control across all conditions. In addition, the method had higher power of both DIF-g and DIF-s for those conditions that had large enough sample sizes to generate enough numbers of masters and non-masters. For the conditions with either larger sample size of masters or non-masters, the power was higher for DIF-s or DIF-g, respectively. MH using total score as matching criterion yielded inflated Type I errors for scenarios with simulated DAF, and both MH methods had consistently poor power in detecting non-uniform DIF.

The results of the real data example indicated the model-based method for DAF detection was capable of identifying group weakness and strength conditioning on general ability. However, model-based DIF detection in the real data didn't show as much difference compared to the two MH methods as was found in the simulation study. This likely occurred because the simulation study suggested the model-based method performed better than the two MH methods only when non-uniform DIF existed or when the attribute correlations were low. Neither of these occurred in the real data example. In addition, high guessing parameter estimates in the real data suggested either the Q-matrix was not completely accurate for the FCAT test analyzed or the test itself did not conform well to the Q-matrix. This is an unfortunate but common condition in real data when a test has not been developed to conform to a specific Q-matrix. Although this reality currently limits the application of cognitive diagnostic based DIF detection, we believe the method developed in this study should become increasingly useful as cognitive diagnostic models becomes more commonly used in test construction.

5.2 Limitations and Future Studies

First, Type I errors for DAF in some of the simulation conditions seemed to lack control. There were no clear patterns of inflated or deflated error rates over the simulated conditions. The relatively small numbers of replications may be one reason for the lack of Type I error control. More replications might improve the estimation of Type I error and power of DAF.

Second, the model developed in this study tried to estimate all possible DAF and DIF simultaneously in one step, thus it required estimating all item parameters and attribute parameters differently across groups. In this way, more parameters were estimated than might have been useful. The subsequent model may not have been the best fit to the data. Several steps could be taken to reduce the model. First, several nested models could be constructed besides the full model. As an example, one could compare a model assuming all parameters to be the same for both groups and a model assuming only attribute parameters varied across groups or a model assuming only item parameters varied across groups. Bayesian model fit criteria might then be used to select the best fitting model.

Third, this study only examined if DAF or DIF significantly differed from 0. In real data, many of the values for these were negligible although they might have also been statistically significant. In the future, it might be useful to have an effect size measure for DAF, DIF-g and DIF-s that could be used jointly with a significance test. The effect size of DIF-g and DIF-s illustrated in the real data was the same as that used as the effect size for MH methods. The value for the effect size using this measure was then compared to the criterion established for MH method. It might be more useful to establish an effect size measure and criterion specific to these parameters.

Finally, even though the model based method in this study had some advantages in DIF detection over the other two MH methods, some easy-to-use, non-model-based methods are still worth exploring. For example, the MH methods used in the paper were all standard MH methods, not capable of detecting non-uniform DIF. At this point, a non-uniform MH might be tried in the same sense that the traditional uniform MH method was extended by Mazor et al. (1994) for non-uniform DIF detection. In addition, the number of attribute mastery patterns increase exponentially with an increase in the number of attributes, and the numbers of examinees for some patterns could be sparse. This was the case in the real data example. Matching on the basis of attribute mastery pattern, therefore, will be not effective in such a case, and the Type I errors likely could be inflated. Since the Q-matrix and the attribute mastery pattern for a given examinee jointly determine whether the examinee has mastered all required attributes for an item, examinees can be simply classified into masters or non-masters at the item level. Matching at the item level would be reduced to two groups for each item, and the problem of large numbers of groups and sparseness for some patterns could be reduced. In this way, however, the matching criterion will be less informative with only two matching groups. The alternative is using the profile of item-required attributes as matching criterion, thus, the number of matching groups can also be reduced without losing too much information.

Bibliography

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

[2] Ackerman, T. A., & Evans, J. A. (1993, April). *A didactic example of conditioning on the complete latent ability space when performing DIF analyses.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

[3] Ackerman, T. A., & Evans, J. A. (1994). The influence of conditioning scores in performing DIF analysis. *Applied Psychological Measurement, 18*, 329-342.

[4] Ackerman, T. A. (1994). Creating a test information profile for a two-dimensional latent space. *Applied psychological Measurement, 18*, 257-275.

[5] Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

[6] Chaimongkol, S . (2005). *Modeling differential item functioning (DIF) using multilevel logistic regression models: A Bayesian perspective.* Unpublished doctoral dissertation, The Florida State University.

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

[8] Cohen, A.S., Kim, S.-H., & Baker, F.B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*, 335-350.

[9] de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333-353.

[10] DiBello, L. V., Stout,W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment liklihood-based classification techniques. In P. D. Nichols, D. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-389). Hillsdale, NJ: Lawrence Erlbaum.

[11] Doignon, J.-P., & Falmange, J.-C. (1999). *Knowledge spaces.* New York: Springer-Verlag.

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

[13] Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland, & H. Wainer (Eds.), *Differential item*

functioning (pp. 35-66). Hillsdale, NJ: Lawrence Earlbaum.

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

[15] Dorans, N. J. & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. Em R. E. Bennett W. C. Ward (Orgs.), *Construction versus multiple choice items in cognitive measurement* (pp. 137-166). New Jersey: Lawrence Erlbaum.

[16] Douglas, J. A., Roussos, L. A., & Stout W. (1996). Item-bundle DIF hypothesis testing: identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*, 465-484.

[17] Florida Department of Education. (2003, March). *Florida Comprehensive Assessment Test.* Tallahassee, FL: Author.

[18] Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-72.

[19] Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework. *Journal of Educational Measurement, 40*, 281-306.

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

[21] Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* Los Altos, CA: Peninsula.

[22] Hambleton, R. K. & Rogers, J.H. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*, 313-334.

[23] Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

[24] Haertel, E. H. (1989). Using restricted latent class models to map the attribute structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.

[25] Henson, R. A., Templin, J. L., & Willse, J. T. (2007, April). *Defining a family of cognitive diagnosis models using Log-linear models with latent variables.* Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.

[26] Holland, P. W. (1985). *On the study of differential item performance without IRT.* Paper presented at the meeting of the Military Testing Association, San Diego.

[27] Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

[28] Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272.

[29] Li, F., Cohen, A. S., Kim, S. H., & Cho, S. J. (in press). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement.*

[30] Li, H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*, 647-677.

[31] Lord, F. M. (1980).*Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

[32] Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179-193.

[33] Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 33*, 379-416.

[34] Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological*

Measurement, 52, 443-45 1.

[35] Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22*, 357-367.

[36] Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*, 131-144.

[37] McGlohen, M., Chang, H.& Miller, E. (2004, April). *Joining diagnostic assessment with large-scale standardized testing.* Paper presented at Annual Meeting of American Educational Research Association, San Diego,CA.

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

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

[40] Milewski, G. B., & Baron, P. A. (2002, April). *Extending DIF methods to inform aggregate report on cognitive skills.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, Louisiana.

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

[42] Narayanon, P., & Swaminathan H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257 - 274.

[43] Pine, S. M. (1977). Applications of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing*: *Proceedings of a symposium presented at the 18th annual convention of the Military Testing Association* (Research Rep. No. 77-1, pp. 37-43). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

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

[45] Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207.

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

[47] Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

[48] Roussos L., & Stout W.(1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

[49] Samuelsen, K. (2005) *Examining differential item functioning from a latent class perspective.* Ph.D. dissertation, University of Maryland, College Park.

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

[51] Smith, B.(2005). *BOA: Bayesian output analysis program user manual* (Version 1.1). [Computer Software]. The University of Iowa, *http://www.public-health.uiowa.edu/boa.*

[52] Spiegelhalter, D., Thomas, A. & Best, N. (2003). *WinBUGS* (version 1.4) [computer program]. Robinson Way, Cambridge CB2 2SR, UK: MRC Biostatistics Unit, Institute of Public Health.

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

[54] Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 215-231). Mahwah, NJ: Erlbaum.

[55] Tatsuoka, K. K. (1995). Architecture of knowlege structures and cognitive diagnosis: A statistical pattern recognition and classification approach. Chapter 14 in Nichols, P. D., Chipman, S. F. & Brennan, R. L. (Eds.) *Cognitively diagnosis assessment* (pp. 327-359). Hillsdale, NJ: Earlbaum.

[56] Templin, J. (2004). *Generalized Linear Mixed Proficiency Models for Cognitive Diagnosis.* (US Patent Application No. 10,924,069)

[57] Templin, J. L., & Henson, R. A. (2006). Measurement of psychology disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305.

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

[59] Walker, C. M., & Beretvas S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: a cognitive explanation for DIF. *Journal of Educational Measurement, 2001*, 147-163.

[60] Zhang, W. (2007). *Detecting differential item functioning using the DINA model.* Ph.D. dissertation, University of North Carolina at Greensboro.

Appendix A

WinBUGS Code for Modified Higher-order DINA Model

```
# NE: the number of examinees
# NS: the number of skills
# NI: the number of items
# gmem: group membership
# a : attribute discrimination parameter
# beta: attribute difficulty parameter
# gamma: group difference in attribute difficulty
# dt: group difference in the mean of general ability
# theta: examinee's general ability parameter
# g: guessing parameter
# s: slip parameter
# alpha: mastery status for each attribute
# q: Q-matrix entry

 model
{
 # Higher-level for DAF Detection

for (j in 1:NE){
for (k in 1:NS){
         logit(pi[j,k])<- a* (theta[j]+dt*(gmem[j]-1))-beta[k]-gamma[k]*(gmem[j]-1)
         alpha[j,k]~dbern(pi[j,k])}

         theta[j]~dnorm(0,1)
         a.theta1[j]<-theta[j]-mean(beta[])/a }

         dt~dnorm(0,1)
         a.dt<-dt-mean(gamma[])/a

         a~dnorm(0,1)I(0,)

for(k in 1:NS){
beta[k]~dnorm(0,1)
a.beta[k]<-beta[k]-mean(beta[])
gamma[k]~dnorm(0,1)
a.gamma[k]<- gamma[k]-mean(gamma[]) }


# Lower-level for DIF Detection

 for (j in 1:NE) {
      for ( i in 1:NI) {
          for (k in 1:NS) {
               x[i,j,k]<-pow(alpha[j,k],q[i,k])
}
      eta[i,j]<-x[i,j,1]*x[i,j,2]*x[i,j,3]*x[i,j,4]*x[i,j,5]
      p[i,j]<-pow(1-s[gmem[j],i],eta[i,j])*pow(g[gmem[j],i],1-eta[i,j])
      r[j,i]~dbern(p[i,j])
} }

for(i in 1:NI){
g[1,i]~dbeta(Ug,Sg)
g[2,i]~dbeta(Ug,Sg)
s[1,i]~dbeta(Us,Ss)
s[2,i]~dbeta(Us,Ss)
difg[i]<-g[2,i]-g[1,i]
difs[i]<-s[2,i]-s[1,i] }

Ug~dunif(.1,.9)
Sg~dunif(.5,10)
Us~dunif(.1,.9)
Ss~dunif(.5,10) }


list(NE=2000, NI=40, NS=5, q=structure(.Data=c( 1,0,0,0,0,
0,0,0,1,0, ... 0,1,0,0,0, 0,1,0,0,0 ),.Dim=c(40,5)), gmem=c(
1,1,1,1,1,1,1,1,1,1, ... 2,2,2,2,2,2,2,2,2,2), r = structure(.Data
= c(
1,1,1,0,1,1,0,1,0,1,0,0,0,0,0,0,1,1,1,0,0,0,0,1,0,1,0,1,1,1,0,1,0,1,1,1,0,0,1,0,
...
1,1,0,1,1,1,1,1,0,1,1,1,0,1,0,0,1,0,1,1,0,1,1,1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,0,1
),.Dim = c(2000,40)))
```

Appendix B

Convergence figures for two selected conditions

Figure B.1: The trace plots for γ^{adj} for the condition with complex structure, 500 examinees per group, a=6, unmatched ability distribution, DAF and non-uniform DIF

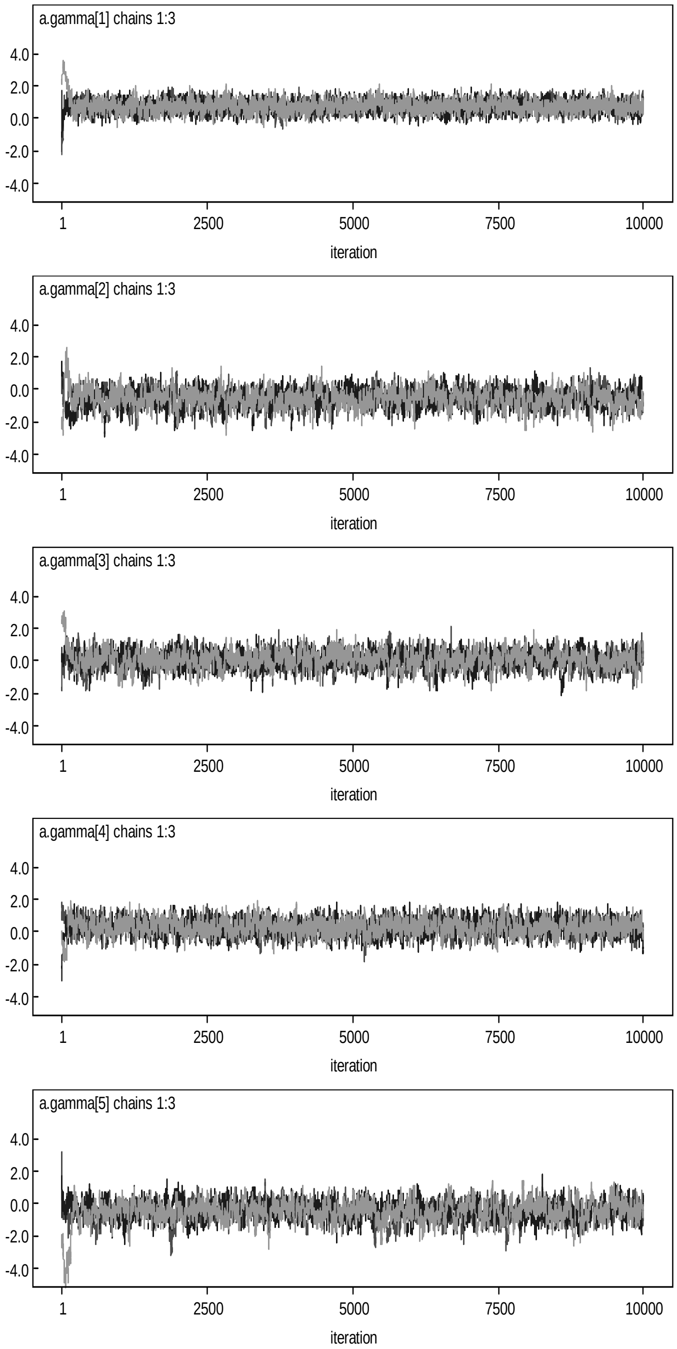


Figure B.2: The plots for Gelman and Rubin Statistic for γ^{adj} for the condition with complex structure, 500 examinees per group, a=6, unmatched ability distribution, DAF and non-uniform DIF

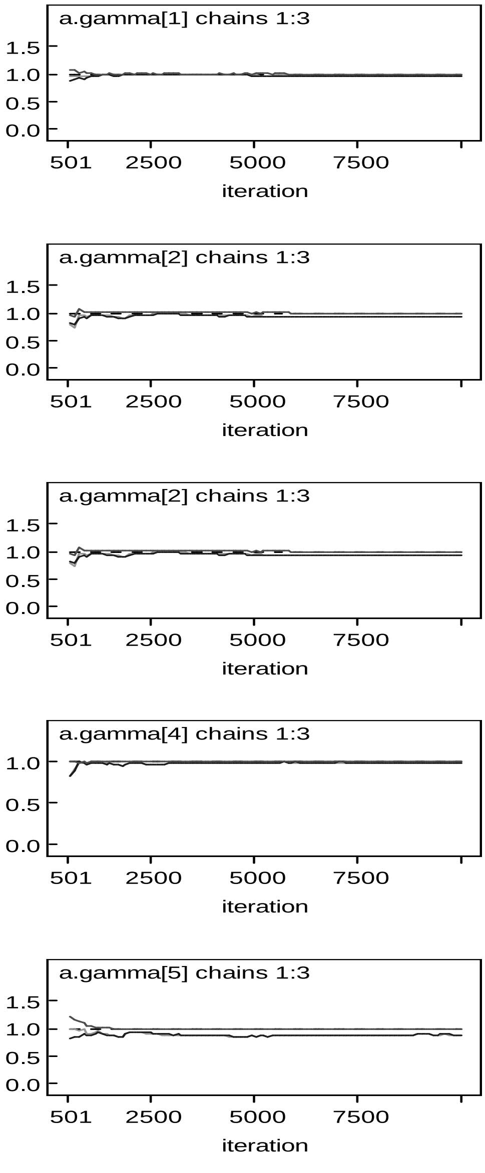


Figure B.3: The trace plots for Δg for the condition with complex structure, 500 examinees per group, a=6, unmatched ability distribution, DAF and non-uniform DIF

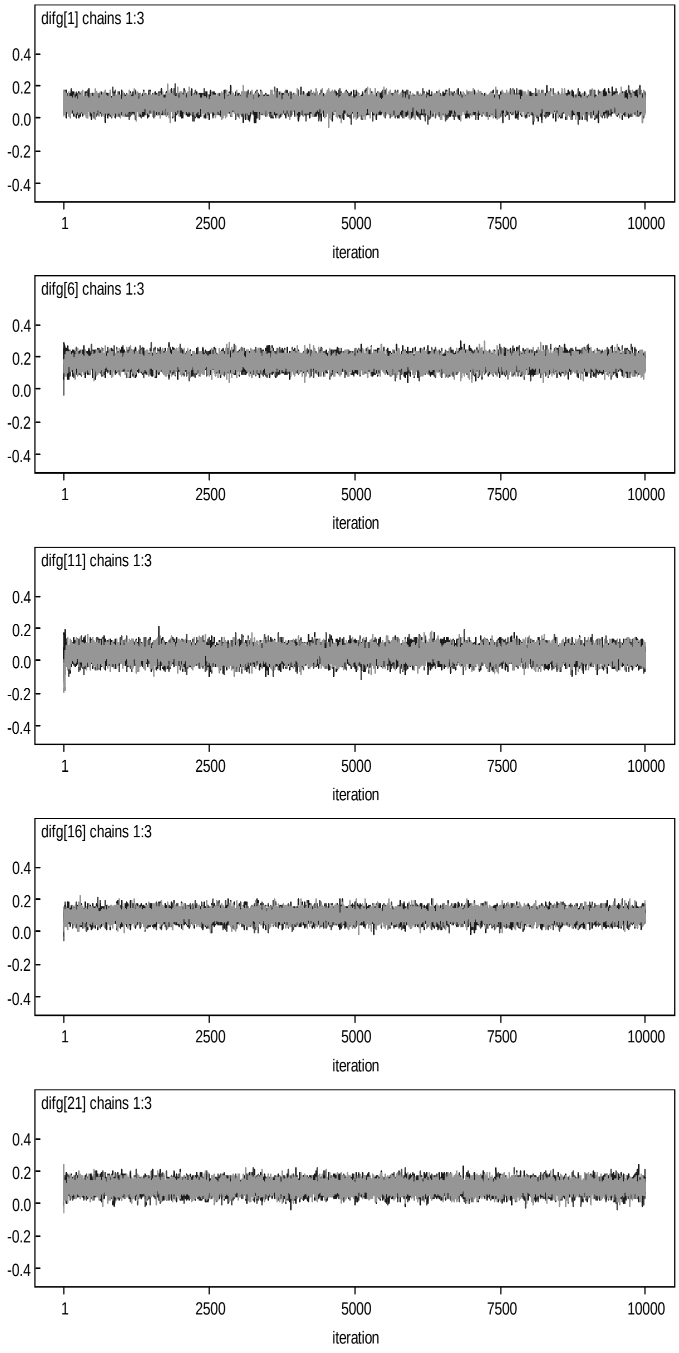


Figure B.4: The plots for Gelman and Rubin Statistic for Δg for the condition with complex structure, 500 examinees per group, a=6, unmatched ability distribution, DAF and non-uniform DIF

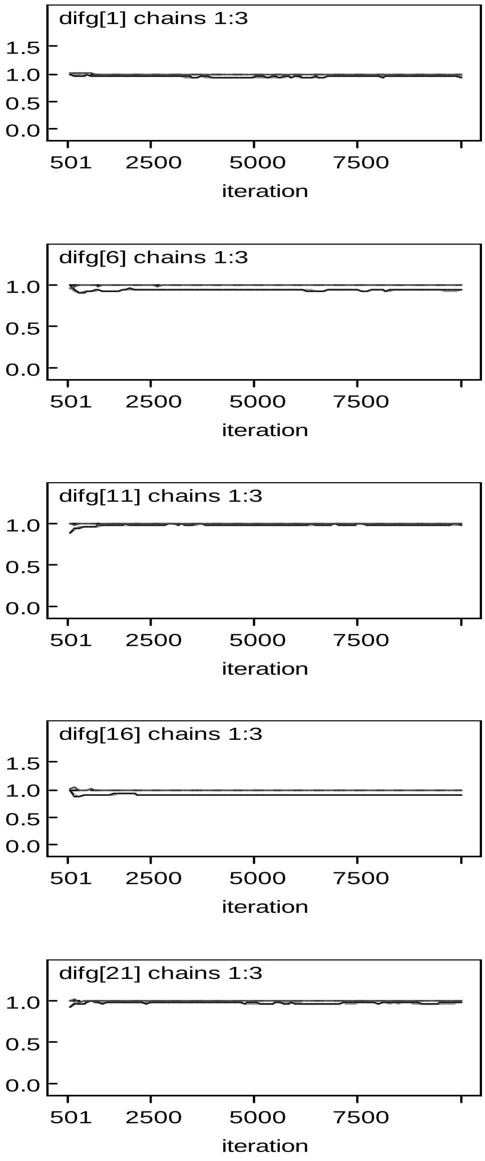


Figure B.5: The trace plots for Δs for the condition with complex structure, 500 examinees per group, a=6, unmatched ability distribution, DAF and non-uniform DIF

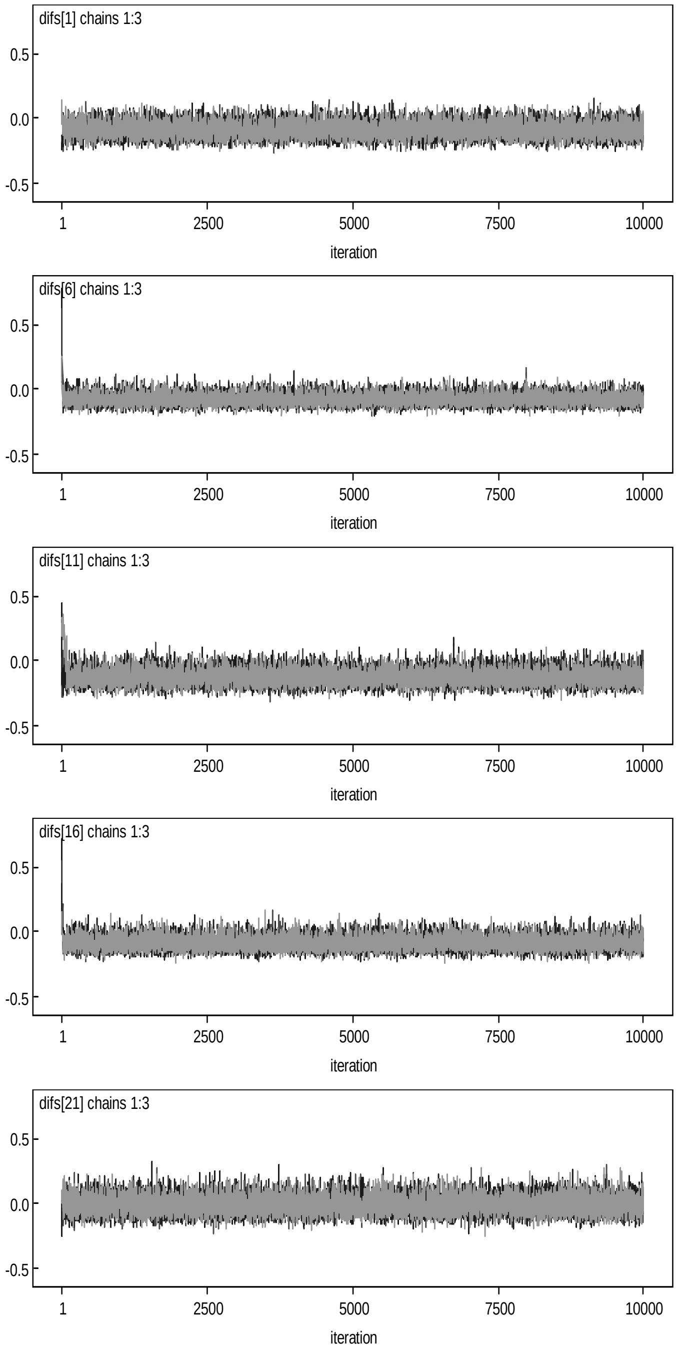


Figure B.6: The plots for Gelman and Rubin Statistic for Δs for the condition with complex structure, 500 examinees per group, a=6, unmatched ability distribution, DAF and non-uniform DIF

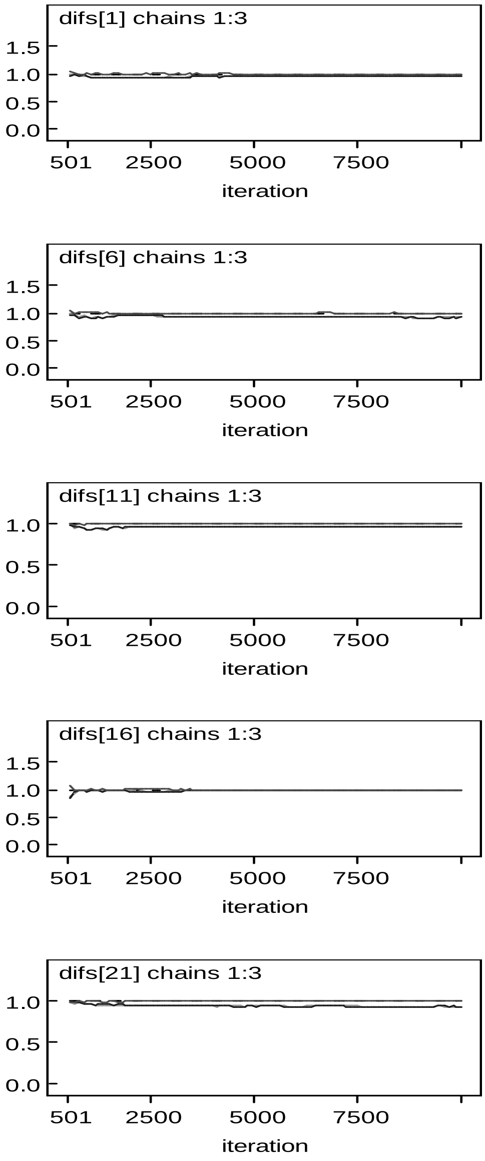


Figure B.7: The trace plots for γ^{adj} for the condition with simplex structure, 1000 examinees per group, a=1, matched ability distribution, DAF and uniform DIF

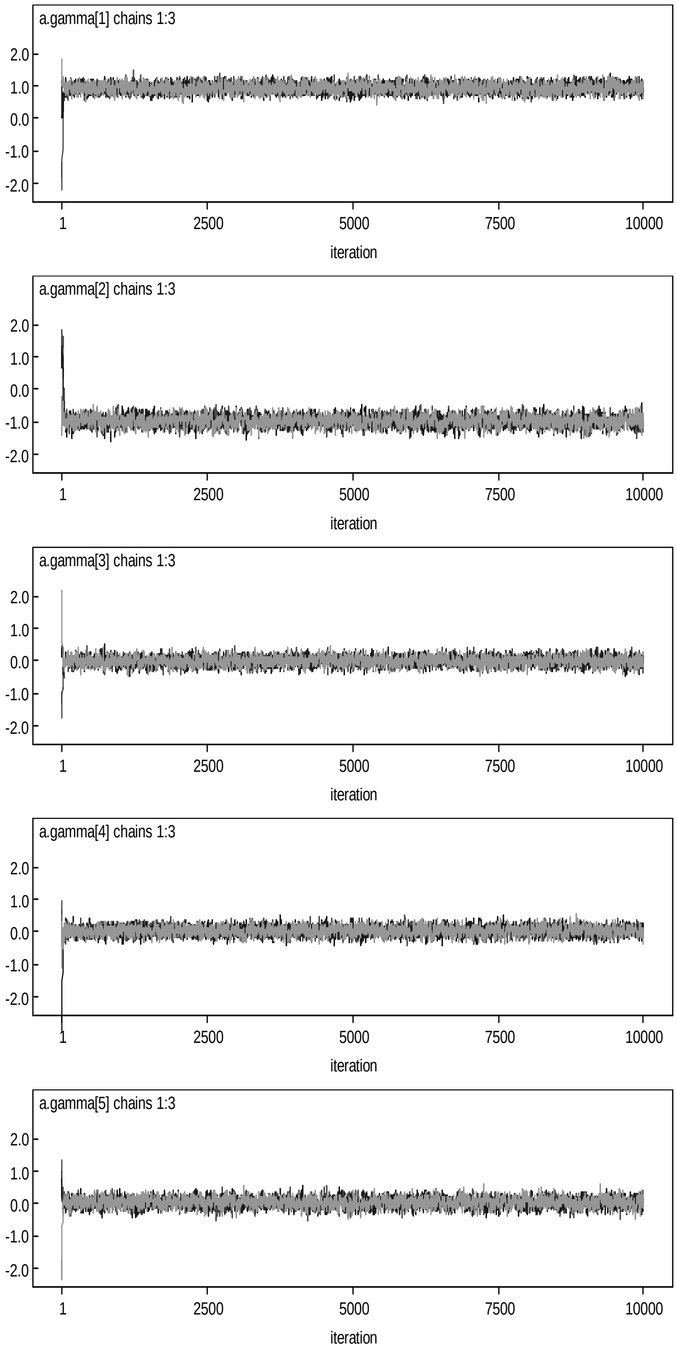


Figure B.8: The plots for Gelman and Rubin Statistic for γ^{adj} for the condition with simplex structure, 1000 examinees per group, a=1, matched ability distribution, DAF and uniform DIF

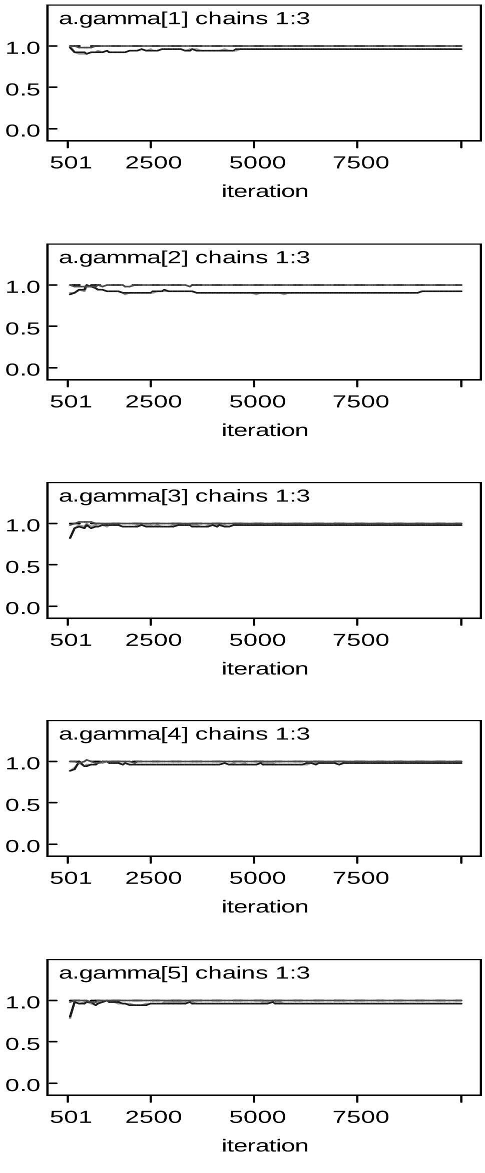


Figure B.9: The trace plots for Δg for the condition with simplex structure, 1000 examinees per group, a=1, matched ability distribution, DAF and uniform DIF

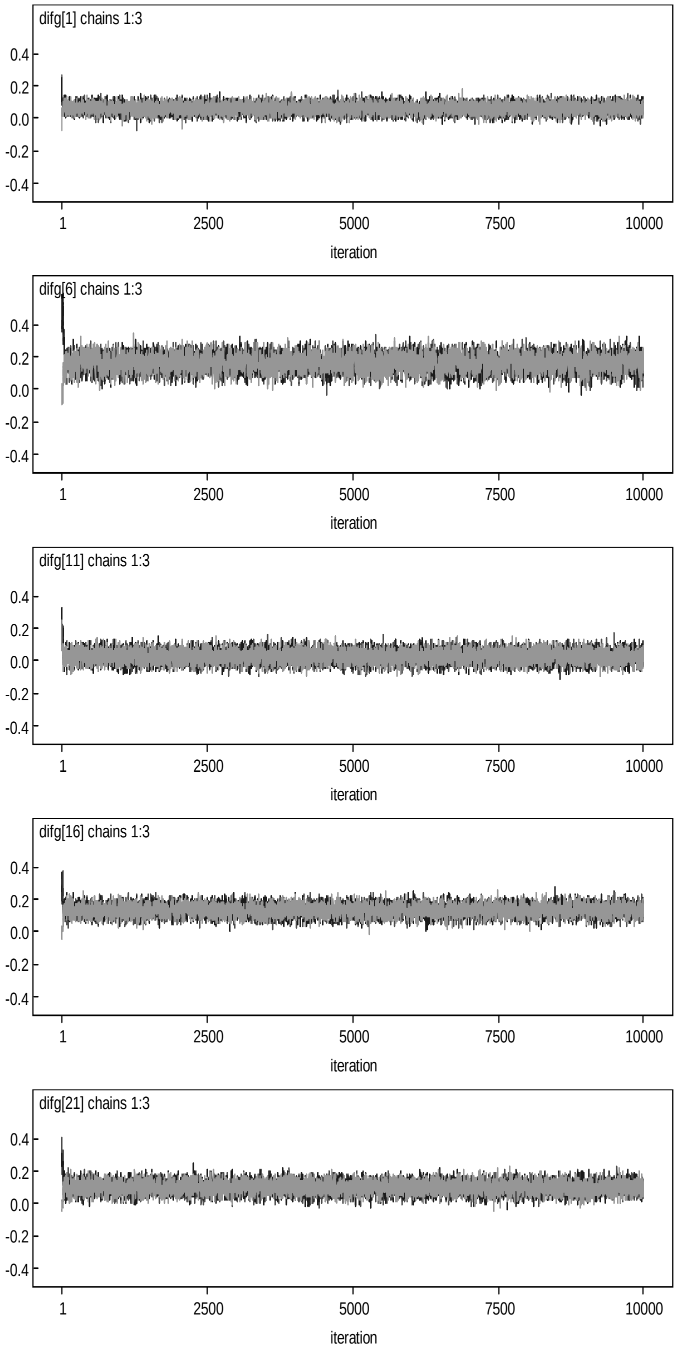


Figure B.10: The plots for Gelman and Rubin Statistic for Δg for the condition with simplex structure, 1000 examinees per group, a=1, matched ability distribution, DAF and uniform DIF

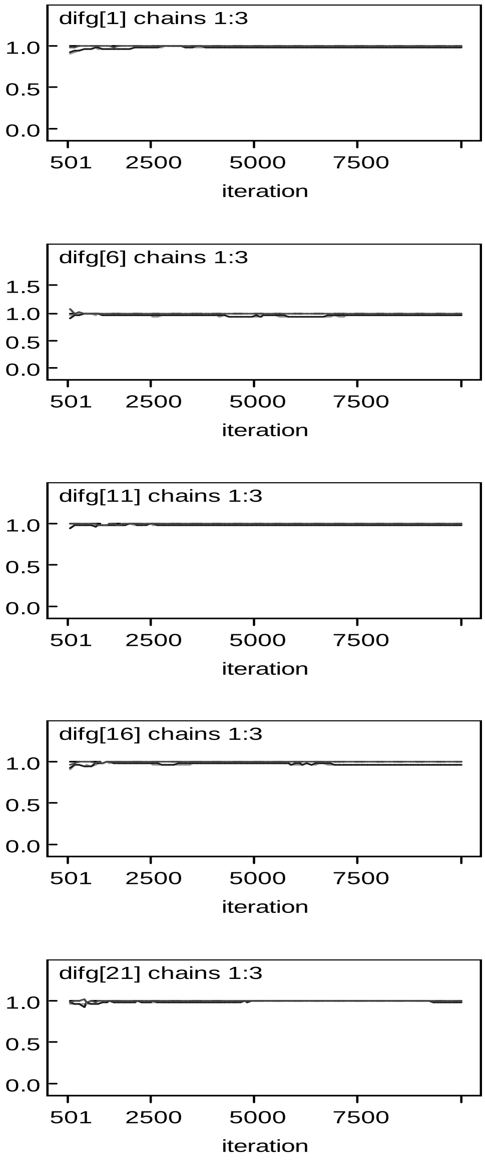


Figure B.11: The trace plots for Δs for the condition with simplex structure, 1000 examinees per group, a=1, matched ability distribution, DAF and uniform DIF

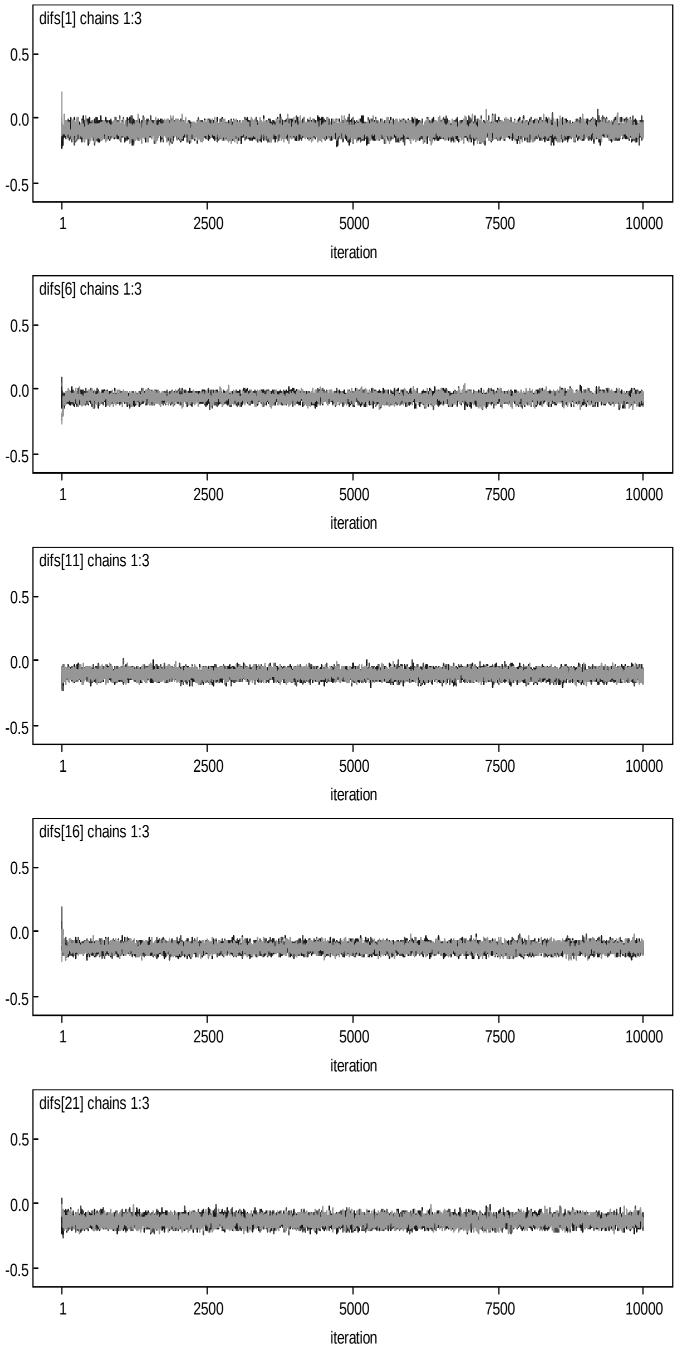
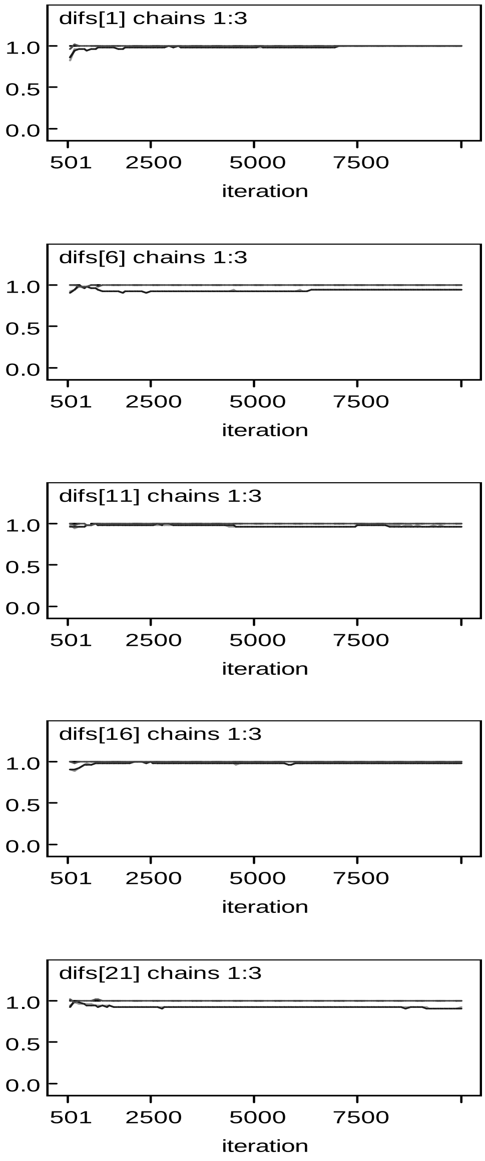


Figure B.12: The plots for Gelman and Rubin Statistic for Δs for the condition with simplex structure, 1000 examinees per group, a=1, matched ability distribution, DAF and uniform DIF


difs[1] chains 1:3
iteration
difs[6] chains 1:3
iteration
difs[11] chains 1:3
iteration
difs[16] chains 1:3
iteration
difs[21] chains 1:3
iteration