ZHANG, WENMIN, Ph.D. Detecting Differential Item Functioning Using the DINA Model. (2006)
Directed by Dr. Terry A. Ackerman and Dr. Robert A. Henson.126 pp.

DIF occurs for an item when one group (the focal group) of examinees is more or less likely to give the correct response to that item when compared to another group (the reference group) after controlling for the primary ability measured in a test. Cognitive assessment models generally deal with a more complex goal than linearly ordering examinees in a low-dimensional Euclidean space. In cognitive diagnostic modeling, ability is no longer represented by the overall test scores or a single continuous ability estimate. Instead, each examinee receives a diagnostic profile indicating mastery or non-mastery of the set of skills required for the test, namely the attribute mastery pattern.

The purpose of the study had three objectives; first to define DIF from a cognitive diagnostic model perspective; second, to identify possible types of DIF occurring in the cognitive diagnostic context introduced into the data simulation design; finally, this study compared traditional matching criteria for DIF procedures, (e.g., total score) to new conditioning variable for DIF detection, namely the attribute mastery patterns or examinee profile scores derived from the DINA model. Two popular DIF detection procedures were used: Mantel-Haenszel procedure (MH) and the Simultaneous Item Bias Test (SIBTEST) based on total test score and profile score matching. Four variables were manipulated in a simulation study: two sample sizes (400 and 800 examinees in each group), five types of DIF introduced by manipulating the item parameters in the DINA model, two levels of DIF amount on a 25-item test (moderate and large DIF), and

three correlations between skill attributes for both groups (no association, medium association and high association).

The simulation study and the real data application demonstrated that, assuming cognitive diagnostic model was correct and the Q-matrix was correctly specified, attribute pattern matching appeared to be more effective than the traditional total test score matching observed by lower Type I error rates and higher power rates under comparable test conditions.

DETECTING DIFFERENTIAL ITEM FUNCTIONING

USING THE DINA MODEL

by

Wenmin Zhang

A Dissertation submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirement of the Degree
Doctor of Philosophy

Greensboro
2006

Approved by

____
Committee Co-Chair

____
Committee Co-Chair

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Co-Chair Dr. Terry A. Ackerman_

Committee Co-Chair Dr. Robert A. Henson__

Committee Members Dr. Richard M. Luecht__

Dr. Scott J. Richter_____

Dr. Jonathan Templin___

October 20, 2006_________
Date of Acceptance by Committee

August 14, 2006__________
Date of Final Oral Examination

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

In the test development, there is an emerging need to design measurement models that allow for appropriate interpretation and use of test data, especially where assessment of higher order thinking is involved. The strong emphasis on measurement models reflects the fact that most theoretical constructs in sociology, psychology, and other social sciences are measured in an indirect way via observations on manifest indicator variables. In this respect, latent variables correspond to the theoretical concepts one tries to measure; manifest variables correspond to the variables that are considered to be indicators of the latent variables. The most common latent variable models include factor analysis, item response models, latent class models, and Bayesian networks. The dependence of item response probabilities on the subject's score on a latent continuum is assumed to obey an explicit functional form, in that the respondent's score and the item parameters play the role of unknown parameters. Those parameters predict the likelihood of all response patterns for each level of competency.

Skills assessment, also referred to as cognitive diagnosis, utilizes latent class models to assess examinee's performance aspects of mental functioning. Cognitive diagnostic models provide additional information that can inform both instruction and learning. One feature of cognitive diagnosis includes its capacity to evaluate a test by assessing the relationship between a set of dichotomous skills and the individual test items, hence improving the qualitative understanding of the latent detailed characteristic underlying an individual's performance. Approaches to cognitive diagnosis serve two purposes; assigning mastery or non-mastery of each skill to each examinee and trying to understand the relationships between the attributes and items.

There has been an increasing pressure to make assessments truly criterion referenced, especially with the No Child Left Behind Act (2001), in which it is mandated that diagnostic reports must be provided to the students, teachers and parents. Those reports should reflect students' achievement related to theory-driven lists of examinee skills, beliefs, and other cognitive features needed to perform tasks in a particular assessment domain.

In the past decades, research was conducted extensively in the field of cognitive assessment (Birbaum, 1968; Junker & Sijstma, 2001; Embretson, 1997; Tasuoka, 1985, 1990,1995; Mislevy, 1996; Maris, 1999, Dibello, Stout & Roussos, 1995). Many statistical models based on a probabilistic approach have been developed to help draw inferences about students' mastery of certain types of knowledge, skills, and strategies to be assessed. The deterministic inputs, noisy "and" gate model (DINA model), for example, assumes examinees must have mastered a set of attributes required by an item in order to answer the item correctly. de la Torre and Douglas (2004) extended the DINA model by proposing a higher-order DINA model which expresses the concept of more general abilities affecting the acquisition of specific knowledge. This model allows the attribute classification and the general aptitude or ability estimation happen at the same time in one consistent model. The introduction of multidimensional latent variable models for cognitive diagnosis has given hope that tests might reveal more information with more diagnostic value than can possibly be reported by using the unidimensional latent trait models (de La Torre & Douglas, 2004). Because of their dichotomous latent classification on each of the examinee ability estimates (mastery or non-mastery), cognitive diagnostic models require new approaches to be developed for assessing and analyzing the validity, reliability, item properties and differential item functioning.

Differential item functioning (DIF) is especially important in test fairness. DIF occurs for an item when one group (the focal group) of examinees is more or less likely to give the correct response to that item when compared to another group (the reference group) after controlling for the primary ability being measured in a test. In Classical Test Theory and Item Response Theory contexts, a variety of techniques, such as Mantel-Haenszel test statistics (MH; Mantel & Haenszel, 1959; Holland & Thayer, 1988), Logistic Regression (LR; Swaminathan & Rogers, 1990), Simultaneous Item Bias Test (SIBTEST; Shealy & Stout,1993) etc., were employed to investigate DIF after conditioning the focal group and the reference group on examinee primary ability that can be represented as the overall test score or the underlying IRT ability estimate (Holland, 1985; Mazor, 1995; Shealy & Stout, 1993).

Standard techniques assessing DIF rely strongly on the unidimensional test structure. When the unidimensional assumption is met, it is reasonable to match examinees on single total test score. Once examinees are conditioned on the total test score, the difference in probabilities of correctly answering an item, when comparing the focal and the reference groups, will be an indication of DIF. The total test score is a sufficient tool to linearly rank and evaluate each examinee's primary ability. One of the shortcomings of using total test score as the matching criterion is the contamination by the inclusion of items containing DIF. In this situation, some researchers suggest a two-step purification procedure where the first step is used to flag DIF items, these items are then removed from the conditioning variable, and the second step uses the purified (i.e., DIF-free) conditioning variable to flag the DIF items (e.g., Clauser, Mazor & Hambleton, 1993; Dorans & Holland, 1993; Holland & Thayer, 1988).

Differential item functioning occurs when the matching criteria do not account for the complete latent space of abilities that was used by the examinees in both groups of interest (Ackerman & Evans, 1994; Clauser, Nungester, Mazor, & Ripkey, 1996). Ackerman (1992) demonstrated how the DIF issue can be eliminated when the complete latent spaces was fully used. When the secondary dimension was present in the test, Clauser et.al (1996) compared the results of the MH procedure and logistic regression for differential item functioning analysis with the matching based on total test score, the matching based on subtest score, and the multivariate matching based on multiple subtest score. In this study when the same matching criteria were used, the MH procedure and logistic regression produced similar outcomes. Of the three different three matching criteria, total test score was the least accurate method. Multiple subtest scores as the conditioning variables were superior to the matching on total test scores and the individual relevant subtest scores (Clauser, Nungester, Mazor & Ripkey, 1996; Mazor, Hambleton, & Clauser, 1998). The conditioning variable on ability has a direct effect on the validity of differential item functioning analysis. When the test structure is not strictly unidimensional, traditional matching criteria using total test scores and one latent ability estimate increases the probability of detecting more items with DIF where actually there are not any (inflated type I error rate) (Oshima & Miller, 1990,1992).

Cognitive assessment models generally deal with a more complex goal than linearly ordering examinees in a low-dimensional Euclidean space. In cognitive diagnostic modeling, ability is no longer represented by the overall test scores or a single continuous ability estimate. Instead, each examinee receives a diagnostic profile indicating mastery or nonmastery of the set of skills required for the test, namely the attribute mastery pattern. Cognitive diagnosis models partition the latent space into more fine-grained, often discrete or dichotomous, cognitive skills or latent attributes, and evaluate the examinee with respect to his/her level of competence of each attribute (Hartz, 2002). In this case, the interpretation and utility of these cognitive diagnosis models resemble or represent a more multidimensional test structure. Thus, the total test score might not be an accurate conditioning variable for investigating DIF from the perspective of the cognitive diagnosis modeling.

Total test score, as a single number, characterizes the proficiency of the examinees in the domain of knowledge. Such scores are not based upon mastery/nonmastery of the underlying skills. An advantage that cognitive diagnostic models have over the general ability estimate for investigating DIF is that examinees' differences in the latent ability are accounted for, to a greater extent, when the conditioning variable is replaced by the skill profiles represented by the skill attribute patterns. The attributes or skills in cognitive diagnosis, relevant to how examinees use knowledge to answer questions, are represented or measured by an individual or a group of items (tasks) in a test.

The purpose of the study has three objectives; first, to define DIF from a cognitive diagnostic model perspective. Second, to identify possible ways of DIF that can occur in the cognitive diagnostic context specified in the data simulation design. Finally, this study compares traditional matching criteria for DIF procedures, (e.g., total test score) to new conditioning variable for DIF detection, namely the attribute mastery patterns or examinee profile scores derived from cognitive diagnostic model through both the simulation study and a real data application. As Sinharay (2004) suggested, no DIF should be redefined such that the focal group and the reference group will have equal success probability for each item after matching on the latent skill mastery profile or classification pattern of the skills.

As in any simulation study, applications of the suggested strategy under a wide range of conditions (different sample sizes, associations between skill attributes, parameter influence in cognitive diagnostics on DIF items, and the amount of DIF introduced, etc.) are examined to establish the degree of generalization of the results obtained. In the end, in order to compare and evaluate the performance of attribute profile score matching for DIF analyses, a dataset from the 1999 Trends in International Math and Science Study (the TIMSS) is used for gender group differential item functioning.

Two popular DIF detection procedures are used: Mantel-Haenszel procedure (MH) (Holland & Thayer, 1988) and the Simultaneous Item Bias Test (SIBTEST)

(Shealy & Stout ,1993). Multiple detection methods are used for this DIF study so that agreement and discrepancy of the outcomes can be compared under various test conditions. Using datasets generated to reflect various conditions of DIF, the Type I error rate and the power rate of the detection procedures are investigated. It is hypothesized that, assuming the cognitive diagnostic model is the correct model, conditioning on the latent attribute mastery patterns will decrease the degree of DIF defined by item parameters differences as compared to the degree of DIF detected using traditional approach of conditioning on the total test score.

CHAPTER II

REVIEW OF LITERATURE

This chapter first reviews the theories and utility of most common cognitive diagnostic models with an emphasis on the Deterministic Inputs, Noisy "And" gate model (the DINA model), and the Noisy Inputs, Deterministic "And" gate model (the NIDA model). Traditional differential item functioning (DIF) detection procedures and their matching criteria are later examined and compared including the Mantel-Haenszel statistic (Holland & Thayer, 1988), and the Simultaneous Item Bias Test called SIBTEST (Shealy & Stout, 1993). Finally, the last chapter discusses the possible DIF applications in the cognitive diagnosis assessment and hypothesizes alternative matching criteria used for cognitive diagnostic purpose.

Cognitive Diagnostic Models

In social science, good models are built to represent a theory. Gulliksen (1961) pointed out that, in test theory, the central issue is focusing on the relationship between an examinee's attribute as measured by the test and the observed scores on that test. In educational and psychological measurement, IRT models have become increasingly popular measurement tools in the past thirty-five years with the strength in estimating an individual's latent ability, based on the information of the person's observed item response vector. IRT methods estimate item parameters and ability parameters on a continuous scale. The item parameters include the discrimination parameter (a-parameter), the difficulty parameter (b-parameter) and the guessing parameter (c-parameter). However, IRT parameters do not provide the information describing what constitutes the difficulty of an item. Often there are situations in practice that one should "look beyond the simple universe of the IRT model – to the content of the item, the structure of the learning area, the pedagogy of the discognitive psychology lines and the psychology of the problem-solving tasks the item demands" (Mislevy, 1993).

In IRT models, an examinee's ability is modeled by θ, the general proficiency parameter. Both θ and the observed scores enable one to summarize, rank, and select examinee's performance in a certain domain. Cognitive diagnostic models make it possible to investigate the mental processes and the content knowledge that underlie the performance by breaking each task down into different elementary components. Rather than reporting a single score, teachers could report student results in terms of a profile, indicating which skills that a student has mastered or not mastered. If students and teachers are aware of the students' skill profile, they are in a better position to know on which skills they need to focus. Strengths and weakness in the learning process can be identified as well as learning intervention strategies. Teachers can then use these intervention strategies to assist students to progress.

For diagnostic purposes, for example, it is useful to know whether an incorrect response in a math test is due to the inability of understanding the information, lack of prerequisite math skills, or the use of an inadequate level of reasoning. An item should be designed in a way that the reason for the incorrect response can be identified. For instance, according to Messick (1984), if a correct response to an item depends on adequate subject knowledge and the possession of certain cognitive abilities, the cognitive abilities should also be assessed, and assessed separately with achievement so that the source of failure in performance can be identified.

The purpose of cognitive diagnostic modeling is to classify examinees into the latent categories based on an array of binary attributes, a vector of latent variables indicating mastery on a set of finite skills under diagnosis. An attribute is identified as a "task, subtask, cognitive process, or skill" involved in the assessment (Tatsuoka, 1995, p.330).

Traditional methods include modeling the rules underlying examinees' responses including deterministic assessment approaches using arithmetic and math data (Birnbaum, 1968). The diagnostic models can be useful in situations where the test is measuring multiple related constructs and where an examinee performance on these constructs is desired. Each item on the test measures these constructs, or cognitive components. The outcomes of the diagnostic models will not focus on the location of examinee's ability in the latent scale, but rather the examinee's performance on each cognitive component. The probabilities can be translated into a profile of the components or attributes that the examinee has mastered. The larger the probabilities on the skills needed to execute the cognitive components of the items, the greater the probability the examinee will get the item right.

Cognitive diagnostic testing utilizes a class of latent attributes (tasks or skills) to identify examinees' mastery level in a set of knowledge. Most current research on cognitive diagnostic modeling focuses on the model selections, examinee parameter fit and item parameters fit evaluation and the identifiability of the parameters (Embretson, 1984; Dibello, Stout & Roussos, 1993; Bolt, 1999; Maris, 1999; Junker, 2000 and Hartz, 2002).

The family of cognitive diagnostic models started with the Linear Logistic Test Model (LLTM) (Fischer, 1973) and Tatsuoka & Tatsuoka's Rule Space model (Tatsuoka & Tatsuoka,1982). Both served as the groundwork on which more elaborate cognitive diagnostic models were developed. The LLTM models how the difficulty parameter of the model is influenced by the cognitive operation by decomposing item difficulty parameters from a logic model into discrete cognitive attribute-based difficulties. In a sense, the LLTM is similar to multidimensional item response models, where the attributes represent more than one dimension. However, the difficulty parameter is not item-specific for each attribute, rather this parameter only indicates the difficulty of an attribute across the whole test.

Tatsuoka & Tatsuoka (1982) developed the Rule Space approach that provided attribute profile scores for each examinee. This approach decomposed examinee ability into cognitive components that could be characterized by a vector of attributes $\boldsymbol{\alpha}$. Since then, many cognitive diagnostic models have been developed and studied. A great contribution by Rule Space approach is the development of the Q-matrix, which establishes the relationship between items and the attributes they are measuring (Tatsuoka, 1990). In a Q-matrix, each row represents an attribute or skill (a vector of $\boldsymbol{\alpha}$) and each column represents a single item. Attributes may include procedures, heuristics, strategies, skills and other knowledge components that are determined by a domain expert. The Q-matrix uses binary numbers (normally 0 and 1) to characterize the attributes required for each item. These binary numbers were first described by Fisher (1973) as the "weight" of attribute k in item i with two possible values 0 and 1. The cell number demonstrates 1 when the knowledge or attribute is required by that item and 0 if the knowledge or attribute is not required by the item.

We can write a $J \times K$ matrix $Q = [q_{jk}]$ of 0's and 1's with entries

$$q_{jk} = \begin{cases} 1, & \textit{if} \quad \textit{attribute} \quad k \quad \textit{is required} \quad \textit{by} \quad \textit{task } j \\ 0, & \qquad\qquad\qquad \textit{if} \quad \textit{not} \end{cases}$$

We will consider the following Q-matrix as representing the relationship between three attributes and a math five-item test. As specified in the Q-matrix, in order to answer the first item correct, the examinee must master attribute one. For item two, attributes one and three have to be both mastered in order to get the item right. Only item three requires the examinee to master all three attributes in order to answer the item correct.

Item	α_1	α_2	α_3
1	1	0	0
2	1	1	0
3	1	1	1
4	0	1	1
5	0	0	1

Q-matrices with attributes determining the item difficulty serves as a bridge that ties the psychometric models with the cognitive processing. The cognitive structure of the test representing the cognitive processes or operations as well their relationship is constructed within Q-matrix. The choice of model and Q-matrix determines the set of equality constraints placed on latent class response probabilities. Current research has found quantified method as well substance evidence to assess the correctness of Q-matrix and to evaluate the efficiency of the Q-matrix that is crucial to provide diagnostic analysis (de la Torre & Douglas, 2004; Henson and Templin, 2006). In most cases, Q-matrix is proposed according to the blueprint standards or according to the judgment of the subject matter experts. It is generally assumed the Q-matrix is a reasonable representation of the latent cognitive structure.

Henson (2004) discussed the quality of the Q-matrix has a direct effect on the estimation of the examinees' attribute patterns. In our example, each attribute is measured by three items and each attribute has a distinctive pattern across items. However, sometimes attributes defined in Q-matrix might be measured by the same items while other attribute patterns are not identified in Q-matrix, thus, examinees with certain attribute patterns couldn't be estimated (Henson, 2004). Simulation studies incorporating a randomly generated Q-matrix thus become a useful and comprehensive tool to investigate the performance of the cognitive diagnostic models with intended cognitive structure specified in the Q-matrix.

Until recently, new advances in psychometric techniques make it possible to evaluate the test by assessing the relationship between the individual skills and the individual test items as well as providing a mastery/nonmastery profile on the attributes or skills measured in the test. The new approaches include the latent class models (Haertel, 1989; Maris, 1999), the unified model (DiBello, Stout & Roussos, 1995), the Reparameterized Unified Model (RUM) with Markov Chain Monte Carlo estimation (Hartz, 2002), multidimensional item response theory (Reckase, 1997), Bayesian networks (Mislevy, 1997), Hybrid model (Yamamoto, 1989) and others. The selection of the approach depends on the purpose of the test, suitability of the model, the identifiability of the parameters and the computing efficiency. For a detailed comparison, see Hartz (2002).

The unified model was developed as an IRT-like model that expressed the stochastic relationship between item response and status of underlying skills (DiBello, Stout and Roussos, 1995). It was based on the Tatsuoka's (1982) rule space model and the latent class response models. The function of unified model is defined as:

$$P(X_i = 1 \middle| \alpha_j, \theta_j) = d_i \prod_{k=1}^{K} \pi_{ik}^{\alpha_{jk} \bullet q_{ik}} r_{ik}^{(1-\alpha_{jk}) \bullet q_{ik}} P_{c_i}(\theta_j) + (1 + d_i) P_{b_i}(\theta_j) \qquad (2.1)$$

where two item parameters are introduced as the "slips" or "guesses".

$\pi_{jk} = P$ (apply skill k correctly in item j | skill k is mastered)

$r_{jk} = P$ (apply skill k correctly in item j | skill k is not mastered)

In addition, c_i and b_i are the guessing and the difficulty parameter from IRT Rasch model not specified otherwise by the Q-matrix, and d_i indicates the probability of selecting the Q-based strategy over all other strategies.

The unified model is the first cognitive model that acknowledges that the Q-matrix is not a complete representation of all the cognitive requirements for the test by building latent attributes outside the Q-matrix and adding additional parameters to improve the fit. The classification reliability proved to be satisfactory in real test-retest data. Unfortunately not all parameters were statistically estimable.

Hartz (2002) reparameterized $\pi_{jk}{}^{*} = \prod_{k=1}^{k} \pi_{jk}$ and $r_{jk}{}^{*} = \frac{r_{j,k}}{\pi_{j,k}}$ and used a Bayesian Markov Chain Monte Carlo (MCMC) framework in programming the estimation software, Arpeggio (Hartz, 2002). $\pi^{*}{}_{jk}$ is a Q-based item conditional difficulty and $r^{*}{}_{ik}$ tells how informative the attribute is represented by item j. Also the alternate strategies have been dropped from the unified model by setting $d_{j} = 1,\ \ i = 1,.....,I$. The reduced the model is defined as

$$P(X_{i} = 1 \middle| \alpha_{j}, \theta_{j}) = \pi^{*} \prod_{k=1}^{K} r^{*(1-\alpha_{jk}) \bullet q_{ik}}_{ik} P_{c_{i}}(\theta_{j}) \qquad (2.2)$$

The Reparameterized Unified Model (RUM) (Hartz, 2002) further reduces the complexity of the parameter space in Unified Model, thus makes the parameters estimable and retains the interpretability of the parameters. The robustness of the RUM item parameters turns out to be satisfactory when Q-matrix is not correctly specified (Hartz, 2002). Items with higher π^{*}'s and lower r^{*}'s provide the most information about the examinees' attribute patterns. r^{*} in the RUM indicates how weak or strong items rely on the attributes in order to discriminate examinees mastery level. If r^{*} is high, the examinees have approximately same probability of getting a correct response regardless of whether the examinees have master the required attributes. Thus, it is likely the attribute is not required by the item, hence, the Q-matrix is reduced by setting corresponding q_{jk} zero. Roussos (1994), Junker (1999) and Hartz (2002) each gave detailed examples and reviews of cognitive diagnostic models that could be used for student assessment purpose. The introduction of multidimensional latent variable models for cognitive diagnosis has given hope that tests might reveal more information with more diagnostic value than can possibly be revealed by a unidimensional latent trait models (de La Torre & Douglas, 2004).

The DINA model

The focus of this thesis is the deterministic inputs, noisy "and" gate (DINA) model. The DINA model was developed by Haertel (1989). He introduced a family of latent class models referred to as binary skills models, under which, examinees are assumed not to possess variable amounts of continuously distributed abilities, but rather to conform to exactly one of a small number of discrete latent classes. The latent class models characterize proficiency in terms of unobservable binary skill attributes, defined independently for a particular set of items. Each examinee's competency can be characterized by of these skills. There is one latent class for each permissible skill pattern. The DINA model divides examinees into a class for lacking all the skills (the null class) and the class for possession of all of the skills (the full class). In other words, an examinee that is missing one of all required attributes is still classified as a nonmaster just as those that haven't mastered any.

The DINA model can be written as

$$P\,[\;Y_{ij}=1\big|\eta_{ij},s_j,g_j\;]=(1-s_j)^{\eta_{ij}}\,g_j^{1-\eta_{ij}} \qquad (2.3)$$

where P denotes the probability of solving the item when examinees possess all of the required skills. η_{ij} is the latent response determined by $\boldsymbol{\alpha}$, the attribute vector for the ith subject and $q_{j,}$, the row of Q-matrix that corresponds to the jth item, can be expressed as

$$\eta_{ij}=\prod\nolimits_{k=1}^{K}\alpha_{ik}^{q_{jk}}. \qquad (2.4)$$

Tatsuoka (1982) defined $\alpha_i=(\alpha_{i1},....,\alpha_{iK})$ as “knowledge states” with $\alpha_{iK}=0$ or 1 depending whether student i possesses attribute k; $\eta_{i.}=(\eta_{i1},.....,\eta_{iJ})$, j =the total number of items, as an indicator of whether all required attributes for each item have been mastered by examinee i, and Y_{ij} as the observed score. For k distinctive attributes, there would be 2^k possible patterns of knowledge mastery.

The relation between a latent variable and the corresponding observed variable is probabilistic and is governed by two classification item parameters unique to each item:

$$s_j=P\big[Y_{ij}=0\big|\eta_{ij}=1\big] \quad \text{and} \qquad (2.5)$$

$$g_j=P\big[Y_{ij}=1\big|\eta_{ij}=0\big], \qquad (2.6)$$

where s_j is the conditional probability of an incorrect response to item j given a latent class as masters (i.e., a false positive probability); and g_j is the conditional probability of a correct response given a latent class as nonmasters (i.e., a true positive probability). s_j can be interpreted as the probability of an examinee answered the item incorrectly or "slip" on the item even though they have mastered all the attributes. When a slip parameter is low, the examinee has a higher probability of answering the item correctly given they have mastered all the required attributes. The parameter g_j can be interpreted as the probability of correct response due to "guessing" even though the examinee has not mastered the required attributes. Maris (1999) alternatively describes g_j as successfully relying on other mental resources. The conditional distribution of the item response variable Y_{ij} also depends on α_{ij} through η_{ij}. Thus, the joint likelihood function of the DINA model, assuming conditional independence as well as independence among subjects, can be written as:

$$L(s,g;\alpha)=\prod_{i=1}^{N}\prod_{j=1}^{J}[s_j^{1-y_{ij}}(1-s_j)^{y_{ij}}]^{\eta_{ij}}[g_j^{y_{ij}}(1-g_j)^{1-y_{ij}}]^{1-\eta_{ij}} \qquad (2.7)$$

Junker (2001) concluded the relationship of the parameters through the calculations for the complete conditional distribution and found that:

1. Estimation of the "slip" probabilities s_j were sensitive only to an examinee's X_{ij} on tasks for which he/she was hypothesized to have all the requisite cognitive attributes (η_{ij}=1)
2. Estimation of the "guessing" probabilities g_j depended only on an examinee's X_{ij} on tasks for which one or more attributes was hypothesized to be missing (η_{ij}=0).
3. Estimation of α_{ij}, indicating possession of attribute k by examinee i, was sensitive only to performance on those tasks for which examinee i was already hypothesized to possess all other requisite cognitive attributes.

The DINA model is a relatively simplistic model, upon which more elaborate cognitive diagnostic models have been developed. de la Torre and Douglas (2004) extended the DINA model to a higher-order DINA model that expresses the concept of more general abilities affecting the acquisition of specific knowledge. This model allows the attribute classification and the general aptitude or ability estimation to occur simultaneously in one model. The formulations of the model have included a higher-order latent trait structure that simplifies the joint distribution of the attributes, and a mechanism for generating the latent responses that accounts for the possibility of multiple strategies (de la Torre and Douglas, 2004). The estimated theta from higher-order DINA model correlated highly with the ability estimates from a two-parameter logistic item response model fitted to the same data. The application of the higher-order DINA model could be fitted to the complex data from The National Assessment of Educational Progress (NAEP), where more intricate sampling designs and mixed test formats were involved.

The NIDA Model

The NIDA model, namely, the Noisy inputs, Deterministic "And" Gate, is another discrete-latent class model that has been the foundation of more complicated cognitive diagnosis model. For example, the Unified Model and the Reparameterized Unified Model (RUM) are extensions of the NIDA model. Unlike DINA model, NIDA provides slip and guessing parameters at the attribute level instead of at item/task level.

The NIDA model is defined as:

$$P\left[\ Y_{ij}=1 \middle| \alpha, s, g\ \right]=\prod_{k=1}^{K}\left[\ (1-s_k)^{\alpha_{ik}}\, g_j^{1-\alpha_{jk}}\ \right]^{Q_{jk}} \tag{2.8}$$

Unlike the DINA model, the NIDA model acknowledges situations where an examinee missing one attribute will have a higher probability of a correct response than those missing all the attributes. The latent response variable in NIDA model is defined as η_{ij}, whether or not student i's performance in the task j is consistent with possessing attribute k. The slip parameter defined for each attribute k is

$$s_k = P\left[\eta_{ijk} = 0 \middle| \alpha_{ik} = 1, Q_{jk} = 1\right] \tag{2.9}$$

and the guessing parameter defined for each attribute k is

$$g_k = P\left[\eta_{ijk} = 1 \middle| \alpha_{ik} = 0, Q_{jk} = 1\right]. \tag{2.10}$$

One extra index defined as the completeness is written as

$$P\left[\eta_{ijk} = 1 \middle| \alpha_{ik} = a, Q_{jk} = 0\right] = 1, \tag{2.11}$$

that indicates the probability examinee masters the attribute k using skills not specified in the Q-matrix, regardless of the value of α_{ij}.

There are a lot of similarities between the DINA and NIDA models. Both the DINA and NIDA models are stochastic conjunctive models for task performance under monotonicity and conditional independence assumptions (Junker, 2000). They are interpreted as the single-strategy cognitive assessment model that posit a stochastic conjunctive relationship between a set of cognitive attributes to be assessed and performance on a particular set of items or tasks in the assessment. All attributes relevant to task performance must be present to maximize probability of correct performance of the task. Each examinee has a discrete proficiency variable associated with their underlying skills. In this latent model, the latent response is 1 if the examinee masters all the required skills or attributes and 0 otherwise. The classification parameters take on two distinct values representing each item mastered or not mastered. Conjunctive models assume that each item requires a set number of attributes that must be mastered in order to respond correctly to the item. Multiple strategies are often accommodated with hierarchical latent class structure that divides the examinee population into latent classes according to strategy. Such an approach uses a different model within each latent strategy class to describe the influence of attributes on task performance within that strategy (Mislevy 1996, Rijkes 1996). As in binary skills latent class models, multiple skills are required for performance. Lacking any of the skills results in lower levels of expected performance, these relationships correspond to conjunctive "AND-gates" in logic.

Markov Chain Monte Carlo Estimation

Applications of efficient algorithms such as maximum likelihood estimation are commonly used in IRT model estimation. However, in the scope of cognitive diagnostic models, new techniques are required because maximum likelihood estimation will produce multiple local maxima when there are a large number of items with more underlying skill dimensions. Markov Chain Monte Carlo (MCMC) algorithms combined with a Bayesian probability framework have become a popular approach to estimation and have been extensively examined by researchers (i.e., Patz & Junker, 1999; Mislevy, 2002, and de La Torre and Douglas, 2004). Because the full conditional distributions of the parameters can not be sampled directly, samples are iteratively drawn from these distributions either using Gibb-sampling or the Metropolis-Hastings algorithm.

Hartz (2002), de La Torre and Douglas (2004) published their MCMC algorithm with Metropolis-Hasting algorithm within the Gibbs sampler to estimate item and examinee parameters for the RUM model and the higher-order DINA model. Inspired by their work, a customized MCMC approach with the Metropolis-Hasting algorithm within the Gibbs sampler was adopted for the DINA model calibration and estimation for the purpose of this thesis. Parameters estimates were based upon the means of the draws of the remaining iterations after burn-in. Bayesian estimation defines a prior probability for each attribute: the probability that a randomly drawn student from the population will have already mastered that attribute or sub-skill. In reality, there exists the situation where one problem can be solved with two or more correct strategies. However, for this study, it will be assumed that items have just one correct solution, that allows them to use a simplified representation, i.e., the Q-matrix (Tatsuoka, 1990), for the relationship between knowledge skills and items.

Although the development of a family of successful cognitive diagnostic models is a great contribution to the cognitive assessment, there is a significant gap in terms of cognitive diagnostic methodology. More work needs to be done in order to flesh out such issues as reliability, differential item functioning, as well as the model's extending applications to different testing scenarios.

Differential Item Functioning

Differential Item Functioning, DIF is especially important to insure test fairness. DIF occurs when one group of examinees has a higher probability of answering the item correctly than another group of examinees after controlling for the valid ability measured by a test (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999).

Commonly used DIF detection procedures include the Mantel-Haenszel statistic (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), item response theory (Raju, 1988; Hambleton, Swaminathan & Rogers, 1991), and the SIBTEST (Shealy & Stout, 1993) procedures. Both Mantel-Haenszel and SIBTEST are nonparametric statistics that do not assume a probabilistic response model. In this study, we will focus on these nonparametric approaches in detecting DIF.

Mantel-Haenszel Method

Mantel- Haenszel test statistics are mostly commonly used to detect differential item functioning (Holland, 1985). Its strength lies in the nonparametric techniques and it can be applied to any dichotomously scored test data. The Mantel-Haenszel method assumes the measures of association within each observed score levels that are similar. This homogeneity assumption allows us to combine strata-specific measures of association to form a single summary measure that has been adjusted for confounding.

Unlike the IRT approach which requires a number of assumptions, an advantage of the MH procedure is the sample size requirement. A sample size of the examinees as small as 200 for the combined group or a minimum of 100 for the focal group is adequate for the DIF detection purpose (Hill, 1989, Kubiak & Kowell, 1990).

The MH procedure requires that these groups be matched according to a relevant stratification. Because there is seldom a clear external factor by which to form the matching criteria, implied levels of ability are used. The ability range of the two groups is divided into k score intervals, and these intervals are used to match samples from each group. A 2x2 contingency table for each of these k ability intervals is constructed from the responses to the suspect item by the examinees of each group:

	Correct on the item	Wrong on the item	Total
Reference group	A_j	B_j	N_{Rj}
Focal group	C_j	D_j	N_{Fj}
Total	$N_{1.j}$	$N_{0.j}$	$N_{..j}$

The Mantel-Haenszel procedure consists of the Mantel-Haenszel odds ratio and the Mantel-Haenszel chi-square statistic. The purpose of this procedure is to compare the odds for success between groups after conditioned on ability.

The Mantel-Haenszel statistic is calculated as:

$$MH = \frac{\left[\left|\sum_{j=1}^{k} A_j - \sum_{j=1}^{k} E(A_j)\right| - 0.5\right]^2}{\sum_{j=1}^{k} Var(A_j)} \qquad (2.12)$$

In which $E(A_j) = \frac{N_{Rj} N_{1.j}}{N_{..j}}$ (2.13)

and $Var(A_j) = \frac{N_{Rj} N_{Fj} N_{1.j} N_{0.j}}{(N_{..j})^2 (N_{..j} - 1)}$. (2.14)

The statistic is distributed as χ^2 distribution with 1 degree of freedom. Odds ratios are used to indicate the difference of correctly answering the item at each of k score levels. The MH statistic can be interpreted as the average amount by which the odds that a reference group member answers an item correctly on an item is larger than the odds for a comparable member of the focal group (Holland & Thayer, 1986). Holland and Thayer (1988) proposed logarithmic transformation of the odds ratio α_{MH} to make the scale symmetric. A value of zero indicates no DIF. A positive number indicates that the item is favoring the focal group and a negative number indicates that the item is favoring the reference group. The logarithmic transformation, sometimes referred as the difference transformation to ETS delta metric, is expressed as:

$$\Delta\alpha_{(MH)} = -2.35\ln(\alpha_{MN}), \tag{2.15}$$

where

$$\alpha_{MH} = \frac{\sum_{j=1}^{k} A_j D_j / N_{..j}}{\sum_{j=1}^{k} B_j C_j / N_{..j}} \tag{2.16}$$

ETS transforms results from the Mantel-Haenszel procedure into its delta units to classify items as one of three types (Dorans & Holland, 1993):

- Negligible DIF, where chi square is not significant and $\Delta_{MH} < 1$

- Intermediate DIF, where chi-square is significant and $1 < \Delta_{MH} < 1.5$
- Large DIF, where chi-square is significant and $\Delta_{MH} \geq 1.5$

These effect size measures were used in the real data application to increase the interpretability of the flagged DIF items.

Simultaneous Item Bias Test Method

The Simultaneous Item Bias Test or SIBTEST procedure approaches the DIF from a different perspective. SIBTEST, developed by Shealy and Stout (1993) is a procedure to detect unidirectional uniform DIF, utilizing a nonlinear regression correction to correct for the inflated Type-I error rate by matching the subtests free from bias. It is a nonparametric approach to detect DIF conditioned on the latent ability. However, the procedure doesn't require or use the IRT ability estimates or parameters for the calculation. A weighted mean difference in item performance, (β_{UNI}), between the focal group and the reference group is computed and then this difference is tested statistically.

The hypothesis for SIBTEST is

$H_0 : \beta_{UNI} = 0$

$H_1 : \beta_{UNI} = 0$

β_{UNI} is defined as

$$\beta_{UNI} = \int B(\theta) f_F(\theta) d\theta, \tag{2.17}$$

in which $B(\theta) = P(\theta, R) - P(\theta, F)$. Conditioned on an estimated latent ability, the difference in the probability of correct response for examinees from the reference and focal groups, $B(\theta)$, is integrated over the ability density function of the focal group. In addition to testing individual DIF items, SIBTEST can also be used to investigate subtests of items that might exhibit DIF. In that case, the subtest item across the ability subgroups is computed as:

$$\hat{\beta}_{UNI} = \sum_{k=0}^{k} p_k d_k \text{ , where } d_k = \bar{Y}_{R_R} - \bar{Y}_{F_k} \tag{2.18}$$

The test statistics for evaluating β_{UNI} is defined as:

$$\beta = \frac{\hat{\beta}_{UNI}}{\hat{\sigma}(\beta_{UNI})} \tag{2.19}$$

$$\hat{\sigma}(\hat{\beta}) = \left[\sum_{k=0}^{k} p_k^2 \left(\frac{\hat{\sigma}^2(Y|k,R)}{N_{RK}} + \frac{\hat{\sigma}^2(Y|k,\mathbf{F})}{N_{FK}}\right)\right]^{1/2} \tag{2.20}$$

Shealy and Stout (1993) demonstrated that SIBTEST has a normal distribution with a mean of 0 and a variance of 1 under the null hypothesis. The null hypothesis is rejected when the SIB statistics exceeds the 100(1-α)/2 from the normal distribution for nondirectional test.

p_k denotes the proportion among the focal group examinees attaining X=k on the valid subtest of items. d_k is the true score mean difference on the studied item for the examinees in the reference and focal groups attaining a subtest score of X=k . $\hat{\sigma}^2(Y|k,R)$ and $\hat{\sigma}^2(Y|k,F)$ are the sample variances of the studied item scores for examinees in the reference group and the focal groups with the same total scores in the matching criteria. However, the test statistic β (defined in 2.19) tends to display inflated Type I error rate caused by the impact-induced group ability difference. Examinees are conditioned on the total test scores are not necessarily matched on their true scores. To correct for the possible bias on the matching criteria, Shealy and Stout (1993) described the procedures for a regression correction on the target ability difference. Each true score is estimated from the observed scores using a linear regression transformation where the slope of the regression equation is the KR-20 reliability for the modified test forms with the studied item deleted from the total test form. In this way, SIBTEST improves upon the observed score matching used in MH.

Like Mantel-Haenszel effect size, guidelines on DIF effect size detected by SIBTEST were proposed by Roussos and Stout (1996b) based on the research findings at Educational Testing Service (Zwick & Ercikan, 1989).

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

In an empirical investigation of DIF, it is recommended that several DIF statistical procedures need to be used for detecting DIF in that the more agreement and consistency that can be found among these procedures, the greater certainty that items detected are items that function differentially (Hambleton & Jones, 1994; Kim & Cohen, 1995; Oshima, McGinty, & Flowers, 1994; Shealy & Stout, 1993). Roussos and Stout (1993) carried out the Type I error rate simulation studies to compare MH and SIBTEST. They demonstrated that the small-sample simulation study showed no large difference in

Type I error performance between MH and SIBTEST although SIBTEST tended to perform slightly superior to the MH. Both the MH and SIBTEST assume asymptotic distributions. Note that MH statistics is distributed as χ^2 with one degree of freedom and SIBTEST is normally distributed $N(0, 1)$. Both procedures turn out efficient when the tests are long or when the variation of discrimination across test items is not large (Roussos and Stout, 1993). However, MH will have potential Type I error problems when the number of examinees is small, and the test is not long, the item response functions are non-Rasch, and/or the reference and focal group observed score distribution display impact (Roussos & Stout, 1996). With impact present, the expected target ability of the reference group tends to be different from that of the focal group, hence causing the inflated DIF detection rate when there is no DIF present. Neither SIBTEST nor MH statistics are sensitive in detecting nonuniform DIF which results from the interaction effect between ability and group membership.

Matching Criteria in DIF

It is a critical step to match examinees on a common measure before their performances are compared. The inclusion of the matching criteria helps to control for the statistical error due to impact or the average ability difference between the focal group and the reference group in the domain the test is measuring. To perform DIF procedure, it is necessary to choose a valid conditioning criterion that can be both internal and external (Clauser, Mazor & Hambleton, 1993). Most of the time, because the standard technique assessing DIF relies strongly on the unidimensional test structure, it is reasonable to match examinees on a single test score that is a sufficient representation of the ability estimate. The most obvious shortcoming of the use of the total test score in the MH procedure as the criterion is the contamination of that criterion produced by the inclusion of items containing DIF, especially when large DIF exists (Clauser, Mazor & Hambleton, 1993). In answer to adjust for this shortcoming, Holland and Thayer (1998) proposed a two-step process in which the MH procedure is implemented first with the total test score as the criterion. Items identified as having DIF are then removed from the conditioning test score, and the MH procedure is re-implemented using this "purified" score as the matching criterion. This process is referred as *purification* of the matching criterion (Clauser et al., 1993; Zumbo, 1999) and *criterion refinement* (Holland & Wainer, 1993). This iterative strategy eliminates the bias when internal criterion, e.g., the total score, is contaminated with a large amount of DIF. The use of matching examinees on this inherently circular internal criterion to carry out DIF analyses is likely to result in less than optimal identification of DIF items and complicate efforts to interpret the findings (Zenisky, Hambleton and Robin, 2003)

Previous research in DIF literature also explored the matching effect of external criteria on DIF detection. These external criteria, that include educational background and experience, were introduced as covariates in addition to the matching test scores to reduce the amount of DIF detected (Kubiak and O'Neill, etc, 1992; Zwick & Ercikan, 1989; Clauser etc, 1996). The results show these additional matching variables helped in reducing the number of DIF items, yet the interpretation should be guided with expert judgment, as these covariate matching criteria might introduce another dimension that is irrelevant to the objective of the test measurement. However, it may provide insights on the factors influencing the performance of test items. Whereas Zwick and Ericikan (1989) also found that the use more rigorous matching criteria did not necessarily reduce the number of DIF items and the MH chi-square statistics differed across the analyses using these different criteria.

In other test situations such as in computerized adaptive tests, the ability estimates in IRT are used as the conditional ability, i.e., in the context of computerized adaptive testing (Zwick, 1993; Nadakuma & Roussos, 2001). Another scenario where number of correct score might not be used as DIF matching criteria occurs in a dimensionally complex test. When test data are multidimensional, matching examinees on the basis of total score is likely to result in a very high percentage of items being identified as DIF even when there are no real differences in the probability of two groups answering the item correct. Roussos and Stout (1996) pointed out if the test produced great impact resulting from the different ability distributions, the total number correct score was no longer a sufficient statistic for the test as the matching trait.

According to Shealy and Stout (1993), the assumption of DIF in IRT context consists of two parts: 1) DIF items elicit at least one secondary dimension, η, in addition to the primary dimension the test is intended to measure, θ, and (2) a difference exists between the two groups of interest in their conditional distributions on the secondary dimension η, given a fixed ability value on the primary dimension, θ . Thus, items that measure the secondary dimension and produce DIF should demonstrate a disproportionate difference between the reference and focal group relative to what should be observed on items that measure only the primary dimension.

Roussos and Stout (1996) interpreted the secondary dimensions further. The secondary dimensions are auxiliary if they are intentionally assessed as part of the construct on the test. DIF caused by auxiliary dimensions is benign. Alternatively, the secondary dimensions are nuisance dimensions if they are unintentionally assessed as part of the construct of the test and hence causing adverse DIF. Douglas, Roussos and Stout (1996) pointed out in order to assess DIF benign and adverse impacts, the matching criterion must result in a construct-valid matching of the examinees on the construct intended to be measured by the test.

Real data and simulation studies have shown the inflated type I error rates when total score was used as the matching criterion for a test that has a nuisance dimension in addition to the primary dimension (Clauser, B. E.,Nungester, R. J., Mazor, K. & Ripkey, D. 1996; Ackerman, T.A. & Evans,1993; Mazor, et al., 1995). Research has also found that the main source of differential item functioning is that the matching criteria does not account for the complete latent space of abilities that was used by the examinees in both groups of interest (Ackerman & Evans, 1993). They demonstrated how DIF issue can be eliminated when two major latent abilities were used. Obviously, in that case, total test score as the matching criteria was not representing the complete underlying latent ability space.

The multidimensionality of the matching criteria becomes an issue in DIF detection. For example, some math tests consist of different type of math items, namely pure algebra or verbal analytical items, e.g., story problems. Several researchers have attempted to match the examinees on subtest scores. Hanzon (1998) used logistic regression technique to match on all the possible traits simultaneously. More extensive research by Mazor, Hambleton, and Clauser (1998) compared the results of MH procedure and logistic regression for differential item functioning analysis with matching based on total test score, matching based on subtest scores and multivariate matching based on multiple subtest scores. Their simulation study involved the variation in dimensional structure, item discrimination parameter and the correlation between traits. When identical matching criteria were used, the MH procedure and logistic regression produced similar outcomes. Logistic regression had the potential advantage over the Mantel-Haenszel statistic for multivariate matching because it avoided the sparseness problems that resulted when examinees were matched on multiple variables in the Mantel-Haenszel procedure. In other cases, logistic regression produced extremely similar results as Mantel-Haenszel procedure (Swaminathan & Rogers, 1990). Within the three matching criteria, total test score was much less accurate than the other two methods and multiple subtest scores simultaneously were superior to matching on total test scores and individual relevant subtest scores. It is the intent of the thesis to explore a matching criterion created from profile scores of mastered/nonmastered skills as determined by a cognitive diagnostic model and compare this approach with the MH and SIBTEST procedures, using a matching criterion of total score approach.

DIF in Cognitive Diagnostic Assessment

Most of the current DIF detection procedures investigate the probability of answering the item correct for focal group and reference group after matching examinees from both groups on an estimate of trait level (total test score or the underlying ability estimate). Cognitive diagnostic models provide a natural profile of scores that can be used as an ability matching variable to investigate DIF. In both the MH and SIBTEST of DIF analysis, examinees were first grouped on the basis of a matching variable that was intended to be a measure of ability in the area of interest. However, for the cognitive diagnostic approach, the conditioning variable was the equivalent of class membership based upon examinee mastered/nonmastered skill profile scores.

Reasons exist why DIF might be equally important in the cognitive diagnostic analysis. First of all, conditioning examinees on comparable ability estimate is a critical step when performing a DIF analysis. The main purpose of cognitive diagnostic analysis is to evaluate the problem examinees have in solving a problem that involves multiple steps of a cognitive process. In other words, the function of cognitive diagnostic models is to classify examinees into mastery and non-mastery groups instead of ranking them. Total score and latent ability, typically used in traditional DIF studies in large scale assessment by testing companies are not meaningful in the cognitive diagnostic context. Second, the methodological and theoretical development in DIF research for the cognitive diagnostic models has not been fully explored. New approaches need to be evaluated to study the impact of DIF for cognitive diagnosis assessment. This study will examinee DIF by matching the examinees on their attribute classification pattern and demonstrate the stability and accuracy of MH and SIBTEST DIF detection procedures compared to traditional total test score.

It is not the intent of the thesis to study the dimensionality assumption check of the cognitive diagnostic model if the Q-matrix is unintentionally mis-specified for the test. It is assumed that the Q-matrix is accurate, when each item is broken down by attribute level and each student will have an estimated profile of mastery that these mastery attribute patterns will provide more information on examinees' ability classifications than a single total score. Based on the mastery of attribute patterns, the classification will result in putting examinees into homogeneous subgroups, that to a greater extent, the impact of group ability difference is removed for DIF study. These attribute patterns indicate student's mastery on a set of skills that are a function of examinee's latent ability.

Sinharay (2004) suggests an elaborate way to calculate a discrepancy measure based on the Mantel-Haenszel test statistic, forming matched groups with respect to their latent skills. In the Bayesian Network context, that is similar to the DINA model in classifying masters when examinees have mastered all the required attributes. No DIF is presented by the same success probability for the focal and reference groups when examinees belong to the same class membership category. Matching with respect to skills (or, attribute profile scores) is a natural concept in this context. The comparison results between regular MH method conditioned on raw score and the posterior predictive MH discrepancy method indicate that reasonable agreement in the detection of DIF, however, the regular MH method tends to show larger p-values, which could lead to more identification of DIF items.

Other researchers have attempted to relate cognitive skills diagnosis to DIF studies from an aggregated level of skills. Milewski and Baron(2002) pointed out DIF detection could be used to report whether the aggregated groups such at school or state levels do better, equally well or worse than the general population based on the notion that differential item functioning at each skill level could be interpreted as differential skill functioning, where a skill is not biased. However, item level difference was replaced by skill level. Strictly speaking, their study investigated the differential skill functioning. Lack of extensive simulation study with no knowledge of the true group difference in cognitive diagnostic analysis prevented them from determining which DIF detection method captured the group differences more accurately.

In keeping up with the development of the DIF detection statistical methods, the interpretations of the cause of DIF using the statistically flagged item remain difficult. To address this problem, Gierl (2004) attempted to use cognitive factors as the organizing principle to provide a substantive basis for generating DIF hypotheses that could be subsequently tested. Gierl (2004) evaluated and studied cognitive skills that elicit group differences based on Roussos and Stout (1996) multidimensionality-based DIF analysis paradigm. A confirmatory perspective was adopted by categorizing subsets of items that share the same cognitive skills principles and then using statistic analysis to confirm males and females performed differently on these subsets. However, Gierl pointed out that DIF statistical methods, in their current form, may not be adequate for testing DIF hypotheses generated from cognitive analyses. There is a need to integrate the statistical DIF analyses with the cognitive theories.

Hypothesis and Research Questions

The purpose of this thesis is to explore possible ways to define DIF for a test by means of the DINA model and investigate the DIF impact under each condition using two DIF detection methods (MH statistics and SIBTEST) with three types of matching criteria (the observed total score, the true score estimate and the attribute profile scores). In addition to demonstrating matching criteria differ, the simulation study helps to determine which matching criterion is superior. A wide range of conditions, such as different sample sizes, parameter influence in cognitive diagnostics on DIF items, attribute associations, and the amount of DIF in the DIF items, etc. are used as factors that influence the detection of DIF defined by manipulating the item parameters in a cognitive diagnostic model.

In cognitive diagnostic models, each item is broken down into the attribute level and each student will have a probability of mastering each skill. Skills listed in the Q-matrix are intentionally assessed as part of the construct on the test. Two distinct groups are formed as a result of the skill analysis, namely those who master the skills and those who do not. DIF impact will be investigated and examined separately for masters and non-masters in the focal and reference groups.

It is hypothesized that, under the assumption that Q-matrix encompasses all the possible underlying skills and that the cognitive diagnostic model is correct, when conditioned on the latent attribute profile scores, the effect of DIF will be more accurately assessed by the MH statistic and SIBTEST than when the total score is used.

In the end, an application of profile score matching criterion of DIF analysis is conducted using a real dataset. The purpose is to demonstrate the capability of performing DIF analysis with profile score matching and to compare its performance with the traditional matching criterion.

CHAPTER III

METHODOLOGY

Differential item functioning (DIF) occurs when an item manifests a different level of difficulty with one group, the focal group, than with another, the reference group. In real test scenario, these groups could be different gender groups or racial groups. This simulation study includes building test conditions into the DINA model and generating new datasets with specific kinds of DIF and a known amount of DIF for both the focal and the reference groups. Next, the attribute mastery patterns for each examinee are estimated using the cognitive diagnostic model. In order to study the performance of the DIF detection procedures under different conditioning variables, the Mantel-Haenszel statistic and SIBTEST are revised to include attribute mastery pattern matching to analyze the examinee response datasets. Type I error rate and power rates are studied to evaluate the performance of each DIF detection procedure.

DIF Simulation Methods

Careful considerations must be given in determining what factors should be built into the simulation study so that the data resemble realistic assessment situations and the conclusions are generalizable. Past research has found that the number of DIF items in simulation studies does influence the validity of the matching variable (Gierl, Gotzmann & Boughton, 2004). Oshima and Miller (1992) simulated conditions with up to 20% of the items exhibiting DIF. Hambleton and Rogers (1989) identified 19% and 25% of the items displaying DIF for a high school proficiency test. A couple of other researchers (Mazor, Kanjee and Clauser, 1993; Raju, Bode and Larsen, 1989) found that, in practical applications, approximately 20% items could show DIF a single test. For a 25 item test, it is reasonable to choose five items with DIF for the simulation study. The first five items are conveniently chosen to have DIF. The total test length is 25 items which is a rather short test. However, cognitive diagnostic models are originally designed to provide feedback on low-stake classroom test to assist in instructions and curriculum interventions, a 25 item test is considered a representing test length where cognitive diagnostic models could be applied.

This simulation study use a 25-item, five-attribute test with randomly generated item parameters and examinee parameters in the DINA model, as described by Junker (1999). In this study, examinee response data are simulated through the same cognitive diagnostic model under a variety of test conditions expected to affect the Type I error rate and power rates of the DIF detection procedures.

Several important factors are carefully considered and selected for the purpose of the simulation study. The generation of each dataset includes the randomly selected Q-matrix, item parameters, examinee parameters, magnitude of DIF for the DIF items and the sample size. Test conditions are simulated based on four factors that are hypothesized to influence DIF detection: two levels of sample size, five levels of item parameter manipulation, two levels of amount of DIF introduced, and three levels of correlations between attributes in examinees response patterns resulting in 60 conditions.

For each study, 25 replications were conducted using the following steps:

1. Randomly generate a Q-matrix
2. Randomly generate the slip and guessing parameters
3. Randomly generate the examinee attribute patterns
4. Simulate examinee responses based on the DINA model
5. Estimate the examinees' attribute patterns using the DINA model
6. Compute the MH statistic and SIBTEST statistic for each generated data sets using

   a) the total test score

   b) the attribute profile score

Although the construction of a Q-matrix is normally based on the expert judgment and previous data analyses, this thesis simulates tests with different Q-matrices to resemble the cognitive structure of a typical test. For the purpose of the thesis, it is assumed that subgroups use the same Q-matrix in the simulation. However, there is a statistical trade-off between the complexity of the Q-matrix and the accuracy of parameter estimation for a test with a fixed length (Hartz, 2002). With few attributes per item, there is insufficient information to estimate the parameters and hence the correct classification rate is reduced. With too many attributes per item, there is insufficient power to differentiate between the attributes. This thesis managed to control the correct classification rates from DINA estimations around 80%.

A 0-1 Q-matrix is generated for a five-attribute 25-item exam with all possible entries, q_{ik} for the Q-matrix. Other specifications constrain the Q-matrix to have one to three attributes measured by each item and any given attribute must be measured by at least four items. This constraint is later used to select the Q-matrix entries. With the generated Q-matrix approximating a reasonable test structure, the possible confounding effect of Q-matrix from causing DIF is removed.

The detection of DIF, in an IRT framework, is normally influenced by the sample size of the subgroups, their ability distribution differences, as well as the magnitude of DIF existing for subgroups of the population using IRT models (Rudas & Zwick, 1997). When cognitive diagnostic models are used, the representations of the group ability distribution differences and parameter differences are changed to the item and examinee parameters differences in the constrained latent class model used for diagnostic purposes. Different probabilities of answering the item correct given the same attribute mastery level should result in DIF between two groups of interest.

The basic problem in the detection of DIF is to differentiate the discrepancies in item difficulties across groups that are due to DIF as opposed to differences in level on the assessed attributes. When examinees are grouped according to the attribute profile scores to control for the ability difference, the item difficulty in the DINA model is represented by the slip and guessing parameters.

The probability for the mastery group to get the item correct is associated with the slip parameter, whereas the probability for the non-mastery group to get the item correct is associated with the guessing parameter. The discrepancies between the slip parameters and the guessing parameters for the focal and the reference groups indicate how informative the item is in differentiating between the masters and non-masters.

For each item, the parameters for the reference group are first simulated, then the parameters for the focal group are varied to simulate different types of DIF and different amounts of DIF that could possibly occur in the process of examinee attribute mastery. By varying the parameters in the DINA model for the focal and reference groups in the simulation study, it is possible to investigate the impact and the power of the DIF detection procedures for masters and nonmasters. For a graphical illustration of different item parameter manipulations to introduce DIF, see the Appendix.

In this simulation, DIF is created in three ways: by changing the slip parameter, by changing the guessing parameter, and by changing both guessing and slip parameter in the focal group. Thus, five distinct types of DIF are examined:

1. As baseline information to compare Type I error rate, both focal group and reference group receive the same set of item parameters. In this way, both groups have equal probability of a correct response for a specific attribute pattern, and hence no DIF should occur for the focal and the reference groups.
2. Increasing only the slip parameter (s) for the first five items for the focal group indicates that the probability of answering the item correctly is lower for the examinees from the focal group who possess all the attributes required by an item when compared to those examinees from the reference group who have mastered those attributes as well. In this way, the p-value of the mastery in the focal group is manipulated to be smaller than the p-value of the masters in the reference group, indicating the item is more difficult for the masters in the focal group. However, because the guessing parameter is unchanged for the focal group, no

DIF will occur between the focal group and reference group for those examinees who have not mastered all required attributes for the item.

3. Increasing only the guessing parameter (g) in the focal group indicates that more examinees use other cognitive strategies in answering the item correctly although they do not possess the necessary attributes required for an item. In this way, the p-value of the non-masters of the focal group is increased, indicating the item is easier for the non-masters in the focal group when compared to the nonmasters of the reference group. However, because the slip parameter is unchanged, no DIF will occur between the focal group and reference group for those examinees who have mastered all required attributes for the item.
4. The slip parameter is increased and the guessing parameter is decreased by an equal amount to produce uniform DIF for the examinees in the focal group. Because the parameter change is only made for the first five items, the p-values for all the examinees in the focal group on the first five items decrease uniformly across masters and non-masters. That is, the item difficulty is increased for the focal group when compared to the examinees in the reference group. The other way to produce the same effect in DIF yet favoring the other group will be to decrease the slip parameter and increase the guessing parameter.
5. Both the slip parameter and the guessing parameter for the focal group are increased resulting in a non-uniform change in the probability of getting the item correct for the examinees in the focal group across masters and non-masters. Specifically, the item is made more difficult for the masters yet easier for the non-

masters in the focal group when compared to examinees of the reference group. This scenario is intended to portray the case in which lower ability population possesses other cognitive strategies in addition to the guessing probability in answering the item correct whereas the high ability group miss the opportunity of demonstrating their ability to answer the item correct.

The guessing parameters, g_j's for the reference group, are randomly generated from, a uniform distribution between .25 and .45, $U(.25,.45)$. The slip parameters, s_j's for the reference group are randomly generated from a uniform distribution between .15 and .25 ,$U(.15,.25)$. Both the upper limits and the lower limits for s_j and g_j are kept small, indicating a discriminating test with more accuracy in estimating examinee's attribute mastery patterns.

After the set of randomly generated slip and guessing parameters are obtained for the reference group, the first five pairs of item parameters for the focal group are manipulated to introduce different types of DIF into the dataset as described in the previous paragraphs. In the No-DIF situation, the focal group examinees receive the same item parameters as the reference group for all 25 items. When parameters are varied to introduce DIF for the first five items, by varying s_j, new sets of s_j's for the first five items in the focal group are created by adding two levels of the amounts of DIF to the old sets of s_j generated from the uniform distribution. In the same manner, new sets of g_j's are changed by means of adding or subtracting two amounts of DIF from the previously uniformly generated g_j according to the test conditions. In this way, the magnitude of the DIF is represented by the amount of variation in the slip and guessing parameters for the first five studied DIF items for the focal group across each test condition. There are two levels of amounts of DIF: a .075 difference between sets of item parameters and a .15 difference. The author believes .075 and .15 sufficiently represent moderate to large DIF as found in the DIF study (Nandakumar & Roussos, 2001).

The 20 items having no DIF in all simulation conditions are used to identify the Type I error rates by calculating proportion of items falsely determined to show DIF. Specifically, both focal and reference groups receive the same item parameters on those 20 items in all simulation conditions, thus No-DIF should exist for these items in a test. Power rates are calculated as the proportion of correctly identified DIF items out all replications on the first five items.

In real educational assessments, attributes specified in the Q-matrix are not necessarily independent with other attributes. Often, the mastery of one of the attribute is dependent upon the mastery of the others. These attributes in cognitive diagnostic models represent the examinees abilities the test is assessing. Simulation study results presented by Henson and Douglas (2005) revealed that the correct attribute pattern classification by cognitive diagnostic models appears to be affected somewhat by the associations between the attributes and the number of attributes. The tetrachoric correlations between attributes are varied in the thesis to indicate the complexity of the ability of the examinees.

A higher positive association between attributes indicates examinees who have mastered an attribute tend to have mastered additional attributes. Likewise, with this high positive association between attributes, examinees who haven't mastered an attribute tend to lack other attributes. Consequently, there could be two dominant types of examinee attribute mastery patterns, either masters or nonmasters of all attributes. In this case, examinee response patterns could be approximated by a unidimensional IRT model where most examinees have either mastered all attributes or not mastered any attributes. Thus, traditional DIF detection procedures that rely on total test scores or latent ability conditioning are likely to perform well in that case. Additionally, decreasing the attribute correlation would be likely to produce more heterogeneous groups of test takers with more categories of attribute mastery patterns, and examinees might exhibit additional dimension of skills in the process. Thus, traditional MH procedure and SIBTEST might not perform well with traditional matching variables. Combined in this DIF study, three levels of attribute pair-wise tetrachoric correlations are established to represent the performance of DIF procedures under different dimensional conditions.

To simulate examinees' attribute patterns, vectors for all examinees are generated from a five-dimensional multivariate normal distribution, $\tilde{\boldsymbol{\alpha}}$, with a mean zero and a 5x5 correlation matrix $\boldsymbol{\rho}$. The next step is to dichotomize each of the generated elements (α_{ik}) in the attribute mastery vectors to either the mastery indicated by one or the nonmastery indicated by zero. Subsequently, a final examinee attribute mastery pattern vector $\boldsymbol{\alpha}$ is created. Zero is then used as the cut-off value to decide the mastery and nonmastery of each attribute for each examinee because the proportion of masters at each attribute is set as 50%, which is equal to zero in the inverse cumulative distribution function of a standard normal probability distribution. Each element in the attribute pattern matrix for the i^{th} examinee and k^{th} attribute, α_{ik}, is defined as

$$\alpha_{ik} = \begin{cases} 1 & if \tilde{\alpha}_{ik} \geq 0 \\ 0 & otherwise. \end{cases}$$

The correlation of the attributes indicates the tetrachoric correlations for all attribute pair-wise associations. There are 5x5 correlation matrices, $\boldsymbol{\rho}$, to specify three levels of associations between α_{ik}'s for the simulation. For the case in which the attributes are considered to be independent of each other, the off-diagonal elements of $\boldsymbol{\rho}$ is zero and the diagonal elements are fixed at one. In the other two cases, all off-diagonal elements of $\boldsymbol{\rho}$ are set to .5 or .8 respectively.

With the attribute patterns and the item parameters, the probability of a correct response for the i^{th} examinee in responding to the j^{th} item, p_{ij}, is calculated using the DINA model (See Equation 2.3). Examinee binary responses $Y_{i.j}$ are defined as

$$Y_{i.j} = \begin{cases} 1 & if \quad p_{ij} \leq \tilde{p}_{ij} \\ & \\ 0 & \quad otherwise. \end{cases}$$

When $\tilde{p}_{ij}$ is randomly generated from a uniformly distribution from zero to one. The sample sizes of the focal group are set to 400 and 800 to represent a moderate and a large sample sizes for DIF study. For the convenience of the study and possible explicit interpretations of the results, the sample sizes are the same for each group. In the real testing situation, it is common for the reference group to have more people than the focal group. Previous research has shown the Type I error rates and power rates for MH, SIBTEST DIF procedures increase as the sample size of the reference and focal groups increase when only significance tests are used (Roussos & Stout, 1996; Jodoin & Gierl, 2001).

To summarize the simulation of the data for all 60 test conditions, the simulation study is conducted by manipulating the item parameters and the examinee ability parameters from the DINA model as well as the sample size for the focal and the reference groups. There are five ways in which the slip and guessing parameters in the DINA model are varied: no item parameter difference between the two subgroups, only the slip parameter is increased, only the guessing parameter is increased, both parameters are changed in the same direction, and both of the parameters are changed in the opposite directions. The amounts of DIF introduced through the manipulation of slip and guessing parameters are set at two levels, .075 and .15. Examinee parameter space is varied by introducing multiple levels of correlations between the attributes with zero being the lowest association, .5 being the medium and .8 being the highest. Datasets are generated using the DINA model for each condition and 25 replications for each test condition are simulated.

DIF Analyses

Under each condition, DIF detection rates are compared to evaluate how reliably DIF detection procedures, namely MH statistics and SIBTEST, perform under specific matching criteria (i.e., the attribute profile scores in cognitive diagnostic model and the total test score).

The DIF analyses are performed using MH statistics and SIBTEST described previously. A significance level of $\alpha = .05$ is chosen for all the analyses for Type I error study.

For each of the 25 replications within 60 conditions, the following calculations are undertaken:

1. MH statistics with the examinee attribute mastery pattern or profile scores as the matching criterion (MH-P)
2. SIBTEST with the examinee attribute mastery pattern or profile scores as the matching criterion (SIBTEST-P)
3. MH statistics with the test score as the matching criterion (MH-T)
4. SIBTEST with the true score estimate as the matching criterion (SIBTEST-T)

The four calculations in each of the 25 replications for each of the 60 conditions yield 1500 individual calibration runs. It should also be noted that the total score and the latent ability estimates are calculated with the studied items included when groups are conditioned on attribute mastery pattern.

When the true parameters are known, the Type I error rate and power of the DIF detection procedures can be accurately assessed. The Type I error rate indicates the probability that DIF is detected for an item when it actually doesn't have DIF. Power rates provide an indication of the probability of correctly detecting the DIF item.

The investigation of the Type I error rate for the DIF detection procedure has practical and significant implications. Any inflated Type I error rates could result in eliminating items which in turn would increase the costs for test development. Previous research has shown highly discriminating items are prone to be falsely identified with DIF, which result in a disproportionate deletion of the informative items (Roussos & Stout, 1996).

To assess the Type I error rates behavior, the 20 items in this 25-item test are modeled to display no DIF in each simulation condition. The Type I error rate could then be investigated by examining the number of No-DIF items mistakenly flagged as DIF items. To study the Type I error rates, item response datasets are generated for both focal and reference groups using the DINA model and the same set of item and examinee parameters, which is the null-DIF case.

The purpose of investigating the Type I error rates is to demonstrate the empirical distributions of the MH and the SIBTEST that are closest to their expected significance levels for the two matching methods. For both the MH statistic and SIBTEST procedures, the empirical Type I error rates are listed for four conditions manipulated in the datasets: a) sample sizes for the focal and reference groups are 400/400 and 800/800, b) examinees with low attribute pair associations to high attribute pair association (tetrachoric correlations for all attribute pairs are equal to 0, .5, and .8), c) the amount of DIF (.075 and .15 difference in slip and guessing parameters between the focal and reference groups), and d) types of DIF represented by parameters manipulation. It is very likely that the Type I error rates for both DIF statistics are at or less than the α level of.05 for the attribute pattern matching method when datasets are generated using DINA model. As sample size increases, the Type I error rate inflation is consistent with the observation that the Type I error rate is greater for larger sample. The Type I error rate is inflated for the total test score matching when the examinees exhibit multidimensional skills in solving the problem because a significant summary value could not account for the examinee difference in the latent knowledge states. Across the sample size and attribute associations, it is also decided to investigate whether the Type I error rates for MH statistic and SIBTEST statistic are consistent with each other with each of the matching criteria.

In addition, the percent of rejections of the studied statistics in the items having DIF yields an empirical estimate of the power of the DIF procedure. In the power study, the 25 pairs of item parameters for the reference group are first simulated, while the first five pairs of item parameters are changed for the focal group to simulate types with two constant amounts of DIF. Power rates are reported based on the first five, indicating how correctly each DIF detection procedure can identify the DIF items with two matching criteria. Cohen (1992) evaluated power rates as excellent if above .80 and moderate if between .80 and .70. It is expected that the correct rejection rate will increase across different sample sizes. The percentage of correctly identified DIF items for the five items is averaged over 25 replications for both the matching criteria. For both the MH statistic and SIBTEST procedures, the power rates are computed and listed for four conditions mentioned previously for the Type I error study. It is anticipated with attribute profile score matching, power rates for SIBTEST and MH are moderate to excellent with consistent results across test conditions than single total score matching. Specifically, the following questions are addressed:

1. Will different types of DIF indicated by the manipulations of item parameters have the same power rates across all conditions? Specifically, will manipulating s_j and g_j in different ways have the same effect in introducing DIF into the items for each condition?
2. Are the power rates consistent across each condition with the two matching criteria for both DIF detection procedures?
3. Will the attribute pattern matching condition result in a higher overall power rate than test score matching?
4. Are the Type I error rates for both DIF detection procedures below the significance level across each condition with the two matching criteria?
5. When the studied-item parameters are manipulated, how do MH and SIBTEST perform with each of the matching criteria under the conditions in which examinee's latent ability attributes are strongly correlated versus independent?
6. What are the effects of sample size and amount of DIF in the DIF simulation analysis?

The primary purpose of the study is to demonstrate that attribute pattern matching is more effective than the traditional total test score matching observed by lowering the Type I error rate and increasing the power rate when datasets with DIF effects are generated in cognitive diagnostic model and how using the correct model can impact our conclusions. In addition, it is equally important to explore cognitive and statistical approaches to define DIF for a cognitive diagnostic test. Another important implication of this study is related to the fact that researchers need not reinvent the existing DIF detection methods to accommodate the cognitive diagnostic based assessment whose purpose is to classify examinees using attribute mastery profile scores.

CHAPTER IV

RESULTS

The results of this study are reported in two parts: Part one summaries the impact of various matching criteria on the Type I error rates for MH statistics and SIBTEST under various test conditions manipulated through sample size, types of DIF, amounts of DIF and the correlations between attributes. Part two reviews the power rates for MH statistics and SIBTEST when conditioning examinees on the true score and the attribute profile score. In each part, comparisons are made on the performance of MH statistics and SIBTEST on the Type I error rate and power rate with both the test score matching and the attribute profile score matching. The performance of each DIF detection procedure is evaluated by examining the Type I error rate for 20 items with No-DIF present and the power study for the five items in which item parameters in the DINA model are manipulated to exhibit DIF across the different test conditions. In addition, an average correct classification rate of 81% was found across replications and test conditions, which ensured correct estimates on the item parameters and the examinee attribute profile score estimates using the DINA model.

Type I Error Study

MH Analysis with Two Matching Criteria

Type I error occurs when an item is identified as having DIF when DIF is not simulated. The Type I error rate was computed as the percentage of detections for all items in all conditions which were not simulated to exhibit DIF. The following results reported the empirical Type I error rate for each of the 60 test conditions that was calculated as the percentage of DIF items for 20 non-DIF items out of 25 replications.

Table 1 and Table 2 illustrate the Type I error study results for MH statistics using total test score matching (MH-T) and attribute profile score conditioning (MH-P) as a function of sample size, DIF types, attribute pair-wise tetrachoric correlations and the amounts of DIF. Table 1 lists the Type I error rates for MH-T and MH-P when the amount of DIF in the DIF-items was equal to a .075 difference in the slip and guessing parameters from the DINA model for the first five items between the focal group and the reference group. Table 2 lists similar output when the amount of DIF was increased to a .15 difference.

Table 1. Type I Error Rates for the 20 Items Averaged over 500 Counts (20x25) When the Amount of DIF =.075. (MH-P and MH-T)

		MH-profile		MH-testscore	
		Group size			
	Correlation	400	800	400	800
No DIF	ρ=0	.040	.048	.054	.056
	ρ=.5	.036	.060	.040	.048
	ρ=.8	.052	.024	.048	.024
Increasing slip	ρ=0	.052	.032	.046	.040
	ρ=.5	.036	.040	.048	.040
	ρ=.8	.048	.026	.040	.040
Increasing guessing	ρ=0	.056	.036	.068	.054
	ρ=.5	.040	.044	.042	.064
	ρ=.8	.030	.042	.042	.046
Increasing both slip and guessing	ρ=0	.046	.036	.044	.062
	ρ=.5	.062	.040	.060	.040
	ρ=.8	.036	.046	.046	.054
Increasing slip and decreasing guessing	ρ=0	.044	.076	.060	.080
	ρ=.5	.044	.056	.072	.106
	ρ=.8	.038	.042	.058	.100

Table 2. Type I Error Rates for the 20 Items Averaged over 500 Counts (20x25) When the Amount of DIF =.15. (MH-P and MH-T)

		MH-profile		MH-testscore	
		Group size			
	Correlation	400	800	400	800
No DIF	ρ=0	.034	.032	.030	.040
	ρ=.5	.048	.030	.038	.042
	ρ=.8	.024	.044	.036	.050
Increasing slip	ρ=0	.036	.046	.032	.052
	ρ=.5	.040	.062	.040	.068
	ρ=.8	.036	.050	.046	.072
Increasing guessing	ρ=0	.052	.070	.106	.174
	ρ=.5	.038	.064	.090	.154
	ρ=.8	.048	.064	.080	.146
Increasing both slip and guessing	ρ=0	.032	.040	.084	.104
	ρ=.5	.048	.050	.082	.080
	ρ=.8	.038	.054	.042	.064
Increasing slip and decreasing guessing	ρ=0	.046	.084	.104	.248
	ρ=.5	.060	.064	.136	.250
	ρ=.8	.050	.074	.126	.250

The results showed the Type I error rates ranged from .024 to .060 across No-DIF test conditions for the two matching criteria. Four out of 24 cases in No- DIF situation yielded the Type I error rates higher than .05 α level, but just less than .06. The low Type I error rates under the significance level in the No-DIF condition suggested attribute profile score matching was a successful and valid empirical criterion for MH statistic in providing the No-DIF baseline information by keeping the slip and guessing parameter the same for subgroups.

With a small amount of DIF, there was no clear pattern in the Type I error rate change as sample size increased within each type of DIF. With a small amount of DIF, when sample size increased from small to medium, eight of the 16 test conditions across attribute correlations and types of DIF yielded a decrease in the Type I error rate for MH-T, and nine out of 16 conditions yielded the Type I error rate decrease for MH-P. When the amount of DIF was larger, there was a similar pattern emerging in the Type I error rate change for MH-T and MH-P. Specifically, when sample size in subgroups was increased, the Type I error rate increased for all attribute correlations and types of DIF conditions. This finding was consistent with the inflated Type I error rate for MH statistic with the larger sample size found in previous research (Ankenmann, Witt & Dunbar, 1999). Table 3 lists the means and the standard deviations of the Type I error rates as a function of sample size and amount of DIF.

Table 3. The Means and Standard Deviations of MH Type I Error Rates as a Function of Sample Size and Amount of DIF

	DIF amount=.075		DIF amount=.15	
	400	800	400	800
MH-profile	.044 (.009)*	.043 (.013)	.051 (.010)	.057 (.015)
MH-testscore	.042 (.010)	.055 (.024)	.071 (.036)	.120 (.079)

*Standard deviations are given within the parentheses.

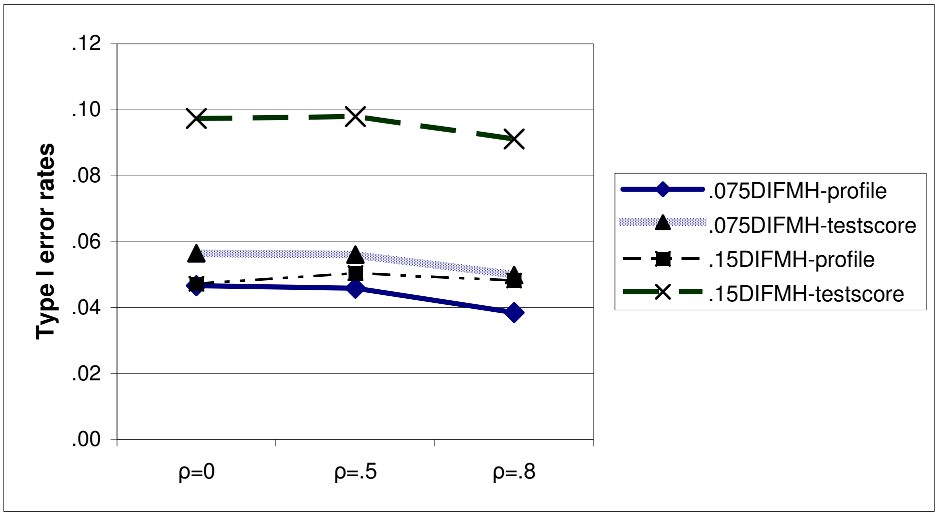
The absolute differences in the Type I error rates between two sample sizes across attribute correlations and types of DIF conditions ranged from .004 to .032 for MH-P and from 0 to .042 for MH-T when DIF amount was small. When the amount of DIF was .15, the differences in the Type I error rates between two sample sizes ranged from .002 to .038 for MH-P and from .002 to .144 for MH-T. It demonstrated that MH-T yielded a larger Type I error rate increase when sample size increased in combination with amount of DIF, whereas the change influenced by sample size with MH-P was stable regardless of the amount of DIF. MH-P was more robust towards the sample size effect with low Type I error rates across test conditions while MH-T yielded some extremely large Type I error rates with a larger sample size and a higher amount of DIF. For instance, with a sample size of 800 in each group and the DIF amount of .15, the Type I error rates for MH-T ranged from .146 to .25 for two DIF type cases, in which the guessing parameter was increased and the uniform DIF case was present.

To evaluate the magnitude of the Type I error rates, the values averaged across amount of DIF were compared with a .05 significance level. When the sample size was

400, MH-P yielded two cases with the Type I error rates above the .05 significance level whereas MH-T showed seven out of 15 cases in which the Type I error rate was higher than the .05 significance level. When the sample size was increased to 800, there were six cases in profile matching and eight in test score matching in which the Type I error rates were larger than the .05 significance level. Overall, MH-P maintained a lower Type I error rate than MH-T across different sample size conditions.

Figure 1 shows that as the amount of DIF was increased, the Type I error rates also increased for MH-T, however, this trend did not have a big effect on the use of MH-P. A larger difference in the change of the Type I error rate occurred for MH-T as compared to using MH-P. Higher amount of DIF almost doubled the MH-T Type I error rates compared to those at a lower amount of DIF. In other words, when larger differences between item parameters from the cognitive diagnostic model for subgroups were created to introduce DIF, total test score matching tended to exhibit inflated Type I error rates. The Type I error rates for MH-P were kept around the .05 significance level.

Figure 1. MH Type I Error Rates as a Function of the Amount of DIF and Attribute Pair-wise Tetrachoric Correlations



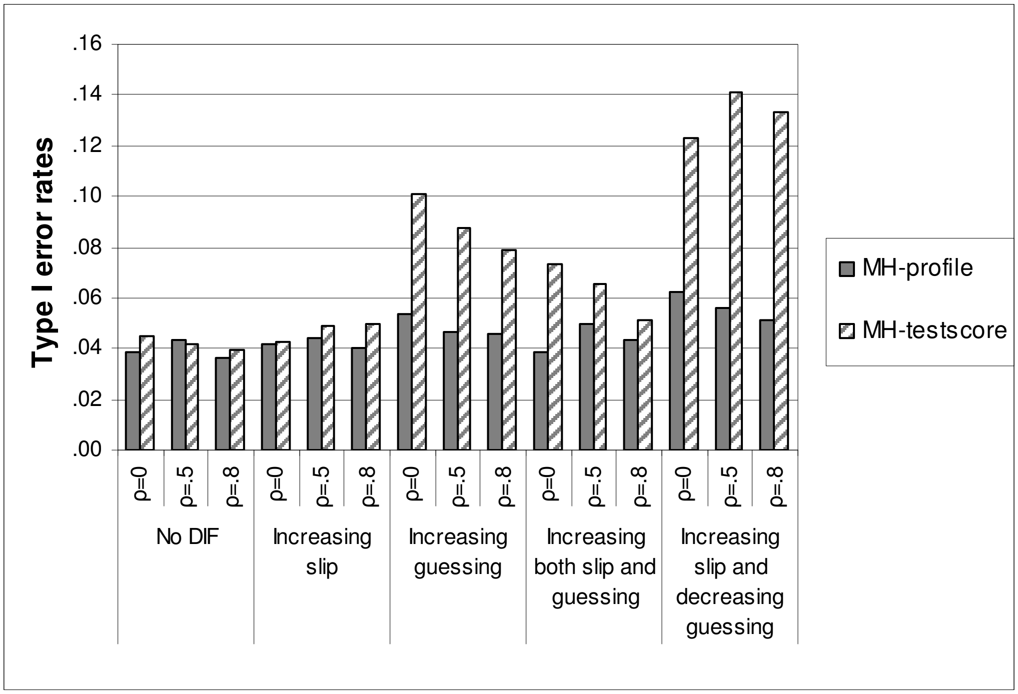
The pair-wise tetrachoric correlations between attributes were set at 0, .5 and .8 to represent examinees' cognitive approach to solve the test questions. Zero pair-wise tetrachoric correlation indicated that examinees were assessed on multiple independent skills that resulted in the testing being multidimensionality. As pair-wise tetrachoric correlations between attribute increased, examinee response data became more unidimensional that can be approximated by a unidimensioanl IRT model because most examinees had either mastered all attributes or not mastered any attributes.

Despite some fluctuations, the Type I error rates for both MH matching methods decreased as the correlation increased in some of the test conditions. This was especially true with a sample size of 400 and small amount of DIF. Figure 1 displays the comparisons on the Type I error rates for different correlations between attributes averaged over types of DIF and sample sizes. As the examinee responses became more unidimensional in high attribute correlation case, the Type I error rates dropped for both the MH statistics for two of the matching criteria and two amounts of DIF. Overall the change in attribute correlations appeared to have had a minimal impact on the Type I error rates for MH-P with a .01 average difference between various correlation simulation conditions.

To investigate the impact of different levels of attribute pair-wise tetrachoric correlations on the Type I error rates as a function of types of DIF, averaged Type I error rates were calculated across two sample sizes and two amounts of DIF and plotted in Figure 2.

A detailed examination on Figure 2 showed the direction of the effect of attribute pair-wise tetrachoric correlations on the Type I error rate was not consistent across the different DIF type conditions. In changing only the guessing parameter and nonuniform DIF cases, as the correlations became higher indicating a more unidimensional item responses, MH-T showed a deceasing Type I error rates. This could be explained by the fact that the total test score did a better job in classifying homogeneous ability groups when the test approximated unidimensional because examinees appeared to be either the masters or the nonmasters of the measured ability. The same decreasing pattern in Type I error rate was observed for MH-P in increasing only the guessing and the uniform DIF cases. Because other DIF types did not produce a significant difference of correct responses between the subgroups due to the nature of the DINA model, attribute tetrachoric correlation effects were not obvious on MH-P Type I error rates. When uniform DIF was present, attribute tetrachoric correlations did not have a clear effect on MH-T because the Type I error rates were extremely inflated.

Figure 2. MH Type I Error Rates Averaged across Sample Size and Amount of DIF



In the No DIF and the increasing the slip parameter cases, the impacts of attribute associations on the Type I error rates were minimal and both MH-T and MH-P yielded similar Type I error rates, each less than the .05 significance level. Larger differences between two MH matching criteria occurred when only the guessing parameter was changed and the slip and guessing parameters are changed in different directions by an equal amount to introduce uniform DIF between the two subgroups. When the slip parameter was increased and the guessing parameter was decreased, uniform DIF was introduced for the focal group as the p-values of correct responses were decreased for both the masters and nonmasters in the focal group. MH-P displayed some much lower Type I error rates than MH-T in both situations. The effect of varying the degree of DIF appeared to be different for MH-P and MH-T as could be observed from the Figure 3 and Figure 4.

Figure 3. Type I Error Rates of MH-P across All Types of DIF

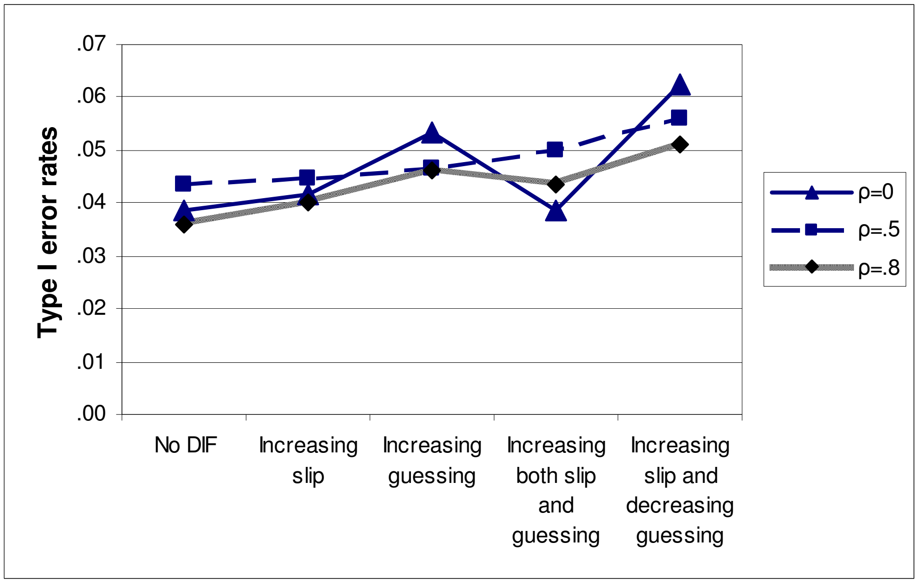
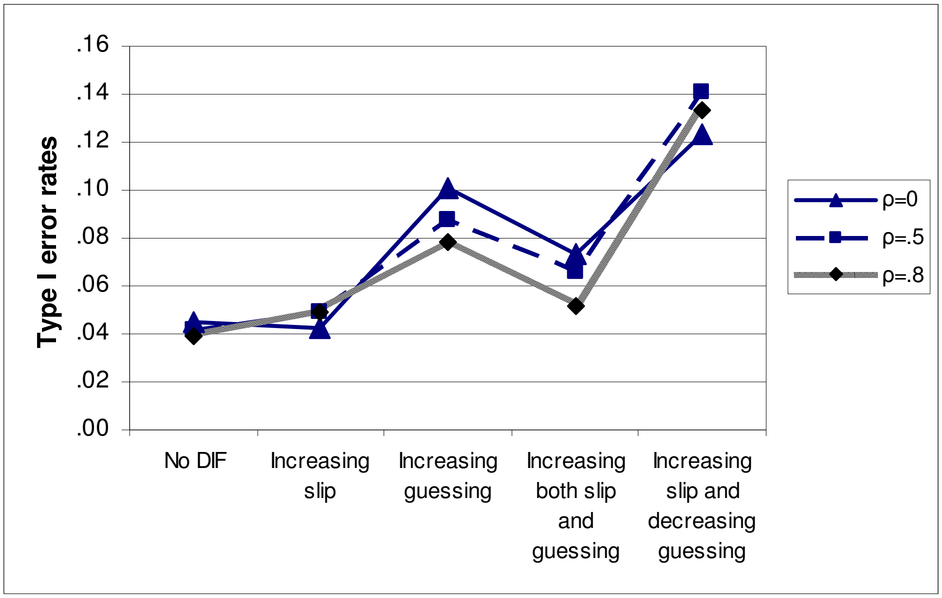


Figure 4. Type I Error Rates of MH-T across All Types of DIF



Both figures depict the Type I error rate for each type of DIF and for each of the attribute pair-wise tetrachoric correlations. The change of the Type I error rates for MH-P was more stable and smaller when compared to MH-T across all types of DIF. When the slip parameter was increased and the guessing parameter was decreased, uniform DIF was introduced for both master and nonmaster groups in focal subgroup. Of the 20 items that did not exhibit DIF, the Type I error rates were the greatest for both matching criteria in the uniform DIF case. Because the datasets were generated using the DINA model, decreasing the probabilities of correct responses for the focal group across master and nonmaster subgroups simultaneously made it harder to classify examinees into the comparable ability levels with the reference group. Thus, items were most likely to be falsely identified with DIF when the additional five items displayed uniform DIF in the same test. However, MH-T seemed to be more influenced with inflated Type I error rates in this DIF case than MH-P.

For MH-T and MH-P, increasing the slip parameter appeared to have the same effect as the no DIF situation with the Type I error rates at or less than the .05 significance level. MH-P appeared to maintain the Type I error rate at the .05 significance level ranging from .036 to .063 across all DIF types whereas MH-T produced much larger Type I error rates than the .05 significance level (ranging from .08 to .14) whenever only the guessing parameter was manipulated or uniform DIF was present.

In summary, the MH-P performed adequately well in classifying people into homogenous ability groups to detect DIF when the cognitive diagnostic model was appropriate. Traditional total score matching for MH statistic appeared to select more heterogeneous examinees to be in the same total test score category especially when larger sample size, higher amounts of DIF and uniform DIF were introduced through the cognitive diagnostic model.

SIBTEST with Two Matching Criteria

Table 4 and Table 5 illustrate the Type I error rate study results for SIBTEST-T and SIBTEST-P as a function of sample sizes, DIF types, attribute pair-wise tetrachoric correlations and amounts of DIF. Table 4 lists the Type I error rates for SIBTEST-T and SIBTEST-P when the amount of DIF in the DIF-items was equal to a .075 difference in the slip and guessing parameters from the DINA model for the first five items between the focal group and the reference group. Table 5 lists similar output only when the amount of DIF was increased to a .15 difference.

Table 4. Type I Error Rates for the 20 Items Averaged over 500 Counts (20x25) When the Amount of DIF =.075. (SIBTEST-P and SIBTEST-T)

		SIB-profile		SIB-truescore	
		Group size			
	Correlation	400	800	400	800
No DIF	ρ=0	.060	.058	.042	.054
	ρ=.5	.048	.070	.048	.054
	ρ=.8	.072	.032	.054	.054
Increasing slip	ρ=0	.068	.040	.044	.046
	ρ=.5	.042	.044	.052	.032
	ρ=.8	.056	.038	.060	.068
Increasing guessing	ρ=0	.060	.046	.072	.042
	ρ=.5	.052	.046	.048	.066
	ρ=.8	.042	.050	.044	.050
Increasing both slip and guessing	ρ=0	.050	.042	.026	.050
	ρ=.5	.066	.044	.064	.052
	ρ=.8	.056	.062	.074	.054
Increasing slip and decreasing guessing	ρ=0	.066	.084	.046	.082
	ρ=.5	.066	.056	.066	.090
	ρ=.8	.058	.052	.046	.064

Table 5. Type I Error Rates for the 20 Items Averaged over 500 Counts (20x25) When the Amount of DIF =.15. (SIBTEST-P and SIBTEST-T)

		SIB-profile		SIB-truescore	
		Group size			
	Correlation	400	800	400	800
No DIF	ρ=0	.034	.032	.046	.070
	ρ=.5	.050	.032	.052	.056
	ρ=.8	.036	.056	.076	.056
	ρ=0	.052	.052	.048	.064
	ρ=.5	.050	.064	.048	.066
Increasing slip	ρ=.8	.042	.062	.062	.052
	ρ=0	.060	.076	.088	.062
	ρ=.5	.042	.068	.068	.090
Increasing guessing	ρ=.8	.068	.068	.076	.098
	ρ=0	.034	.044	.058	.052
Increasing both slip and guessing	ρ=.5	.072	.050	.068	.062
	ρ=.8	.054	.058	.042	.050
	ρ=0	.074	.098	.068	.130
Increasing slip and decreasing guessing	ρ=.5	.076	.084	.078	.122
	ρ=.8	.058	.098	.078	.074

Overall, the SIBTEST based on true score matching (SIBTEST-T) did not keep the Type I error rates under the .05 significance level as frequently as the modified SIBTEST based on attribute pattern matching (SIBTEST-P). The Type I error rates for the SIBTEST-T was .0044 higher than the SIBTEST-P on average across all the test conditions. However, this difference was not considered significant.

It was noted that the regression correction used in Shealy and Stout was not used here because the examinees were matched on the mastery of the latent skills which resembled the true score estimates for examinee ability. In this case, the SIBTEST-P was reduced to Standardized Mean Difference (SMD) technique based on attribute profile score matching.

In the No-DIF situation, SIBTEST-T with a maximum Type I error rate value of .054 appeared to keep the Type I error rates around the .05 significance level more frequently than SIBTEST-P that had the highest Type I error rate of .072 when the amount of DIF was .075. When the amount of DIF was increased to .15, SIBTEST-P had a better control over the Type I error rates with only one case exceeding the significance level when compared to SIBTEST-T that had five cases above the significance level.

The effect of sample size on the Type I error rate of two SIBTEST matching criteria did not seem to follow a clear pattern for all DIF types and attribute pair-wise tetrachoric correlation conditions. Table 6 lists the means and the standard deviations of the Type I error rates averaged across types of DIF and attribute associations.

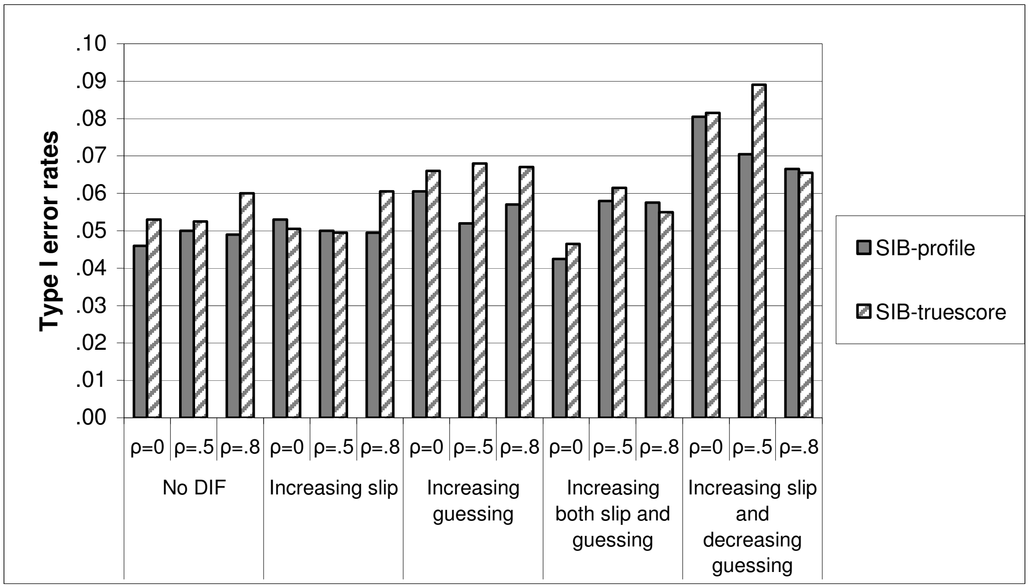
Table 6. The Means and Standard Deviations of SIBTEST Type I Error Rates as a Function of Sample Size and Amount of DIF

	DIF amount=.075		DIF amount=.15	
	400	800	400	800
SIB-profile	.057 (.009)*	.051 (.013)	.053 (.014)	.063 (.020)
SIB-truescore	.052 (.013)	.057 (.015)	.064 (.014)	.074 (.025)

*Standard deviations are within the parentheses.

When the amount of DIF was a .075 difference, SIBTEST-P tended to produce lower Type I error rate when sample size was increased. On the contrary, the Type I error rates for SIBTEST-T tended to increase. With a larger amount of DIF, it was obvious that the Type I error rates for SIBTEST-P and SIBTEST-T increased by .01 on average as sample size increased. This amount of change in Type I error rates was not considered significant as a function of sample size. Table 6 also illustrates that as the amount of DIF was increased, the Type I error rates for both methods increased. When a larger sample size and a higher amount of DIF were both present, the Type I error rate was inflated for both SIBTEST matching methods. Figure 5 demonstrates the Type I error rate as a function of the attribute associations and types of DIF for SIBTEST-P and SIBTEST-T.

Figure 5. SIBTEST Type I Error Rate Averaged across Sample Size and Amount of DIF



The effect of the attribute pair-wise tetrachoric correlations on Type I error rate for SIBTEST-P change was minimal for most test conditions. When the attribute association was constant for the two SIBTEST methods as illustrated in Figure 6 through Figure 8, the Type I error rates for SIBTEST-T were equal to or higher than those for SIBTEST-P with each type of DIF. Both SIBTEST-T and SIBTEST-P maintained stable Type I error rate regardless of the attribute correlation conditions. When the attribute correlation was low indicating a multidimensional item response structure, both method provided similar Type I error rates across five DIF types. As expected, they produced similar results because both methods were built on a multidimensional paradigm by conditioning examinees on the latent multiple abilities. As the correlations became larger in a more unidimensional item response case, the effect of type of DIF started to diminish for both the SIBTEST-T and the SIBTEST-P and the Type I error rate difference between them disappeared for the uniform and nonuniform DIF cases.

Figure 6. Type I Error Rates for SIBTEST-T and SIBTEST-P as a Function of Types of DIF When ρ=0

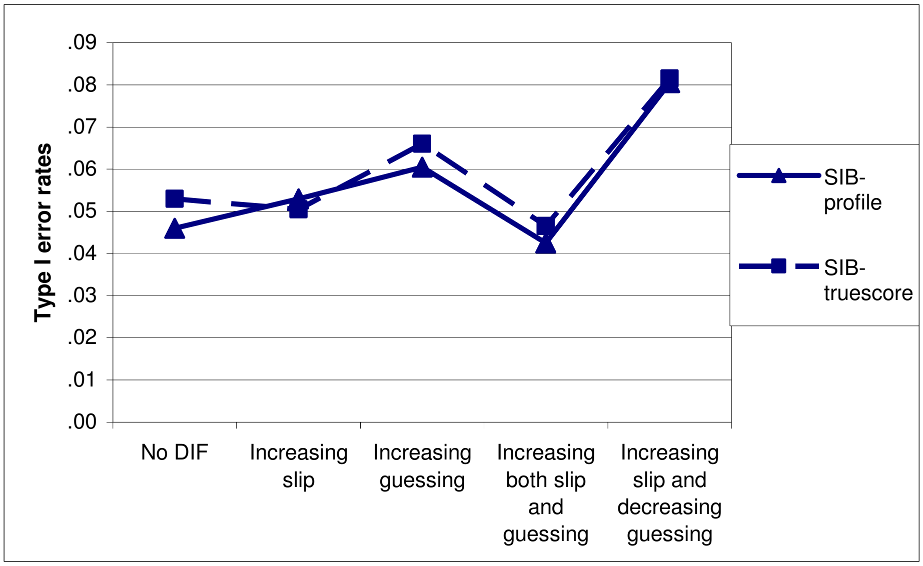


Figure 7. Type I Error Rates for SIBTEST-T and SIBTEST-P as a Function of Types of DIF When ρ=.5

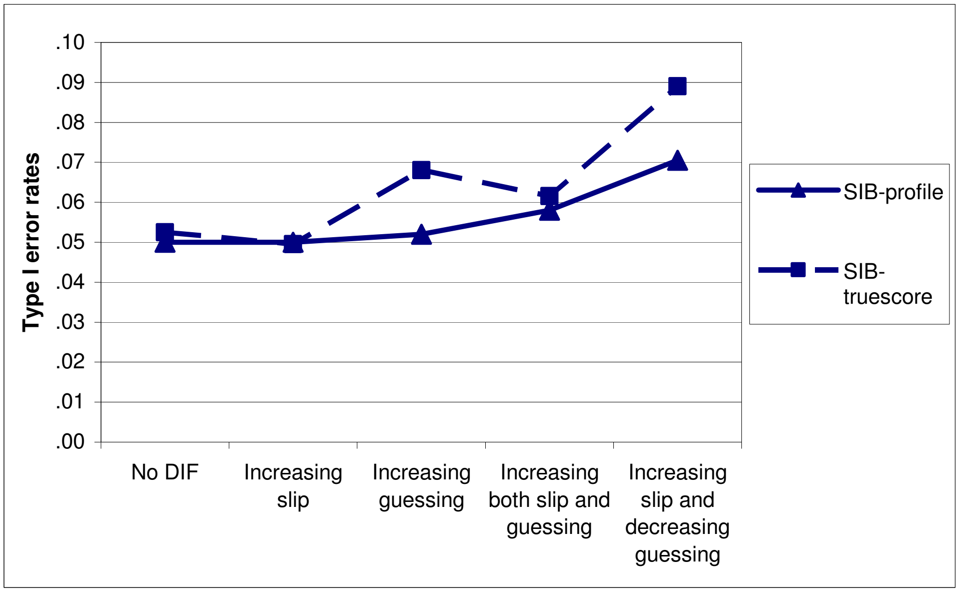


Figure 8. Type I Error Rates for SIBTEST-T and SIBTEST-P as a Function of Types of DIF When ρ=.8

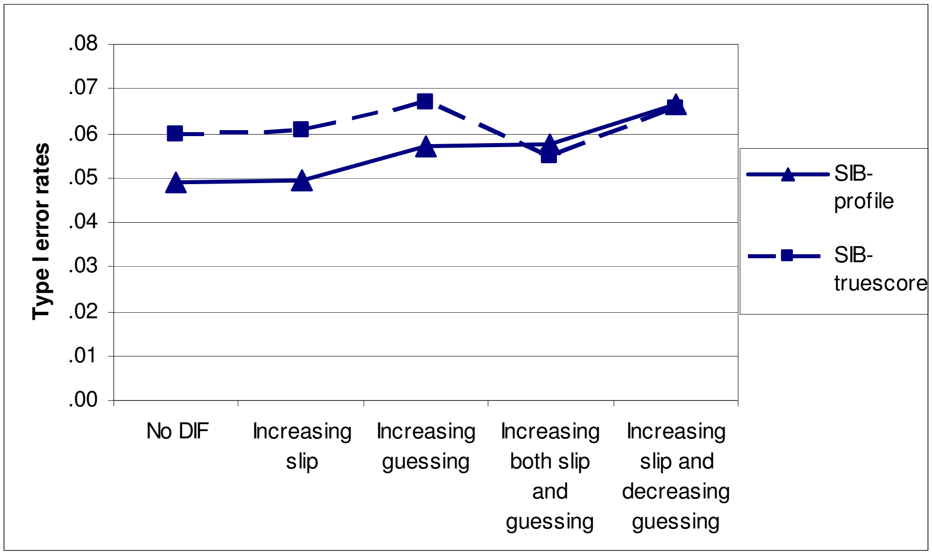


Figure 9 shows that the highest Type I error rate occurred for SIBTEST-P when the slip parameter was increased and the guessing parameter was decreased by an equal amount (uniform DIF). Working in the same fashion, Figure 10 demonstrates that the uniform DIF yielded the highest amount of Type I error rate compared with other types of DIF for SIBTEST-T.

Figure 9. Type I Error Rates of SIBTEST-P as a Function of the Attribute Associations and DIF Types

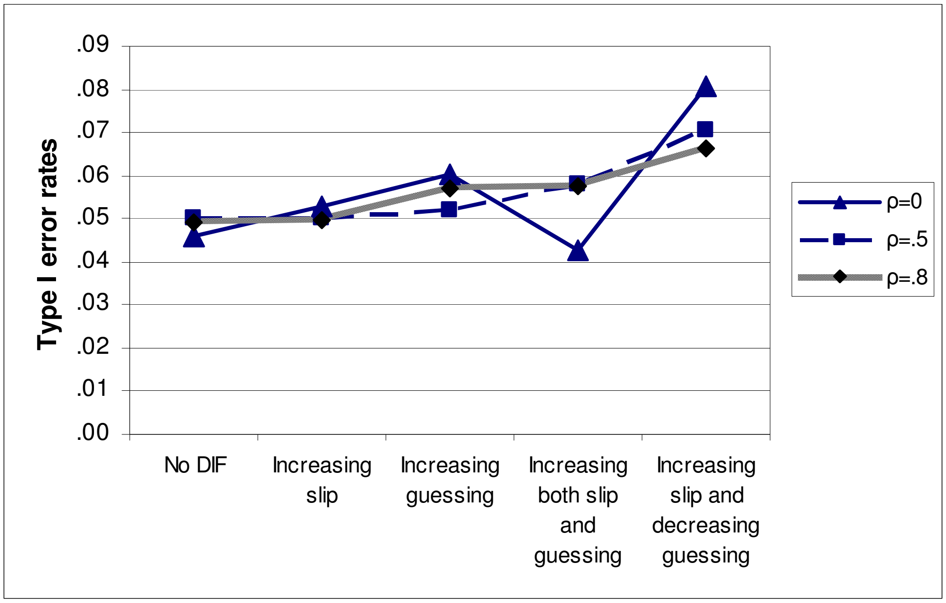
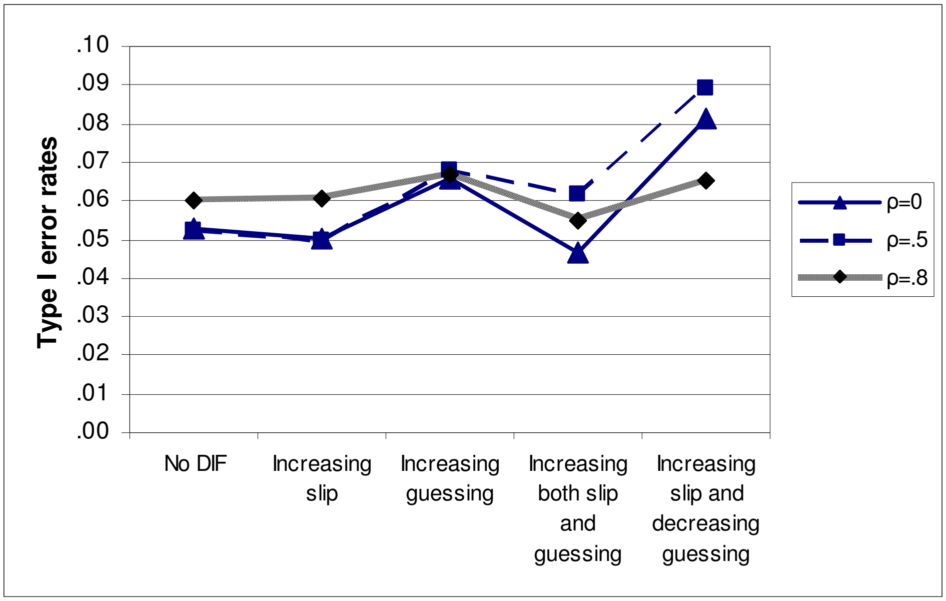


Figure 10. Type I Error Rates of SIBTEST-T as a Function of the Attribute Associations and DIF Types



Consistent with previous research findings that SIBTEST and MH statistics appeared not to be sensitive to the detection of nonuniform DIF. Neither was SIBTEST-P able to detect nonuniform DIF introduced through the DINA model by manipulating the slip and guessing parameters in the same direction for the focal group. Lower Type I error rates were observed in the nonuniform DIF case for both SIBTEST-P and SIBTEST-T when compared with other DIF types. For SIBTEST-T, manipulating the slip parameter yielded similar Type I error rates as in the No-DIF situation. Manipulating the guessing parameter alone resulted in the same Type I error rate for all three attribute association conditions.

Overall, the increase in sample size and amount of DIF resulted in higher Type I error rates for both SIBTEST-P and SIBTEST-T. However, the increase was not significantly large. The effect of the attribute tetrachoric correlations on the Type I error rates of SIBTEST-P and SIBTEST-T did not follow a clear pattern. However, when examinee responses were approximately multidimensional, SIBTEST-P and SIBTEST-T provided similar Type I error rates.

Uniform DIF in combination with a larger sample size and a higher amount of DIF resulted in an inflated Type I error rate, especially for SIBTEST-T. In comparable test conditions, SIBTEST-P yielded a lower Type I error rate than SIBTEST-T indicating that attribute profile score matching appears to be a valid criterion to conduct DIF analysis using SIBTEST technique when the cognitive diagnostic model was the correct model.

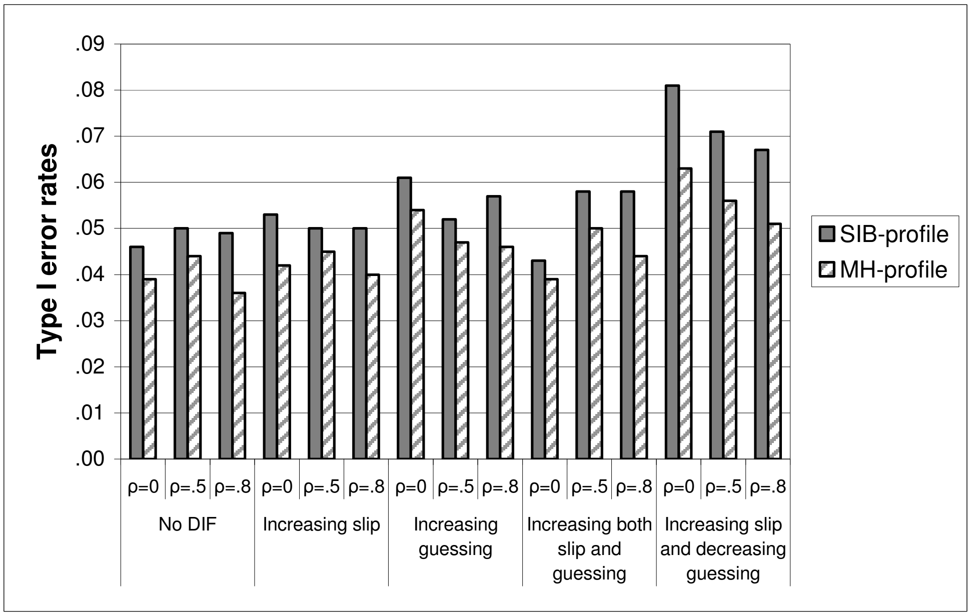
Comparison between MH and SIBTEST

For the purpose of the study, MH chi-square statistic based on profile score matching (MH-P) and the modified SIBTEST with profile score matching (SIBTEST-P) are compared and contrasted in Table 7 and Figure 11.

Table 7. Type I Error Rates Averaged across Sample Size and Amount of DIF (SIBTEST-P and MH-P)

	Correlation	SIBTEST-profile	MH-profile
No DIF	ρ=0	.046	.039
	ρ=.5	.050	.044
	ρ=.8	.049	.036
Increasing slip	ρ=0	.053	.042
	ρ=.5	.050	.045
	ρ=.8	.050	.040
Increasing guessing	ρ=0	.061	.054
	ρ=.5	.052	.047
	ρ=.8	.057	.046
Increasing both slip and guessing	ρ=0	.043	.039
	ρ=.5	.058	.050
	ρ=.8	.058	.044
Increasing slip and decreasing guessing	ρ=0	.081	.063
	ρ=.5	.071	.056
	ρ=.8	.067	.051

Figure 11. Type I Error Rates Averaged across Sample Size and Amount of DIF (SIBTEST-P and MH-P)



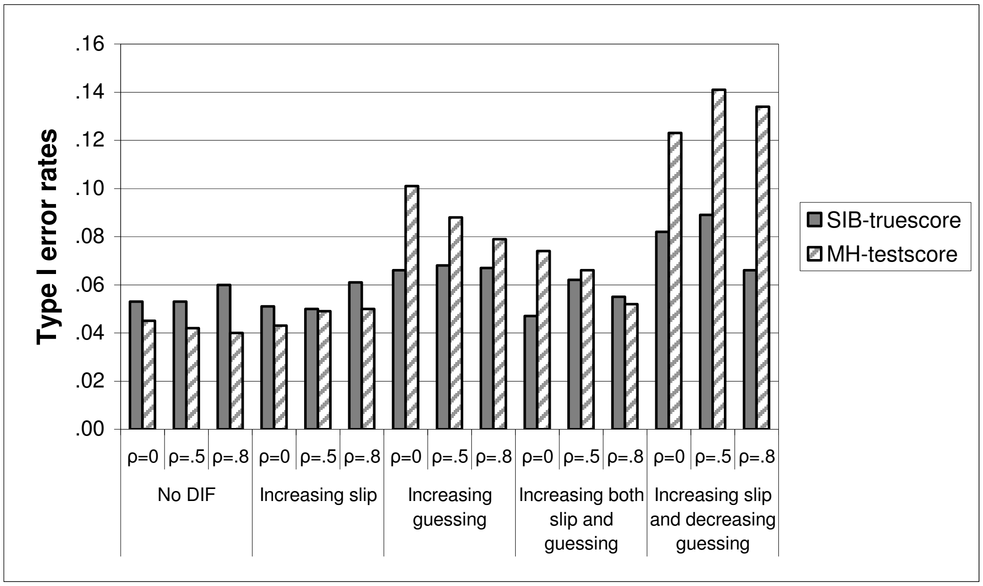
MH-P yielded on average .01 smaller Type I error rates when compared to SIBTEST-P for each type of DIF and attribute pair-wise tetrachoric correlations. The changes of Type I error rates for both MH-P and SIBTEST-P were consistent across three attribute associations embedded with five DIF types. A correlation calculated between Type I error rates for SIBTEST-P and MH-P was as high as .93. When uniform DIF was introduced into the data for the first five items, both methods yielded a Type I error rate ranging from .051 to .08 for the additional 20 items that had no DIF, above the .05 significance level. In addition, the correlations between attributes had minimal impacts on the Type I error rates of the two methods. There was no clear consistent pattern for both methods in terms of the Type I error rate change as a function of the attribute correlation for each type of DIF. However when changing the guessing parameter alone or uniform DIF was present, for both SIBTEST-P and MH-P, as the correlation increased, the Type I error rates decreased. Specifically, with higher attribute pair-wise tetrachoric correlation indicating examinees might manifest dependent skills in their responses, attribute profile score matching was accurate in classifying examinees into homogeneous groups. Hence, DIF statistics were less likely to falsely identify items that contained no DIF when additional items in the test were simulated with uniform DIF. Overall, the MH-P and SIBTEST-P were similar in producing the same amount of Type I error rate across the various test conditions.

MH chi-square statistic(MH-T) and SIBTEST(SIBTEST-T) are compared and contrasted in Table 8 and Figure 12. MH-T yielded on average .013 more in the Type I error rate than the SIBTEST-T for each type of DIF and attribute pair-wise tetrachoric correlations as illustrated in Table 8 and Figure 12

Table 8. Type I Error Rates Averaged across Sample Size and Amount of DIF (SIBTEST-T and MH-T)

	Correlation	SIB-truescore	MH-testscore
No DIF	ρ=0	.053	.045
	ρ=.5	.053	.042
	ρ=.8	.060	.040
Increasing slip	ρ=0	.051	.043
	ρ=.5	.050	.049
	ρ=.8	.061	.050
Increasing guessing	ρ=0	.066	.101
	ρ=.5	.068	.088
	ρ=.8	.067	.079
Increasing both slip and guessing	ρ=0	.047	.074
	ρ=.5	.062	.066
	ρ=.8	.055	.052
Increasing slip and decreasing guessing	ρ=0	.082	.123
	ρ=.5	.089	.141
	ρ=.8	.066	.134

Figure 12. Type I Error Rates Averaged across Sample Size and Amount of DIF (SIBTEST-T and MH-T)



When uniform DIF was introduced into the data for the first five items, both MH-T and SIBTEST-T yielded the Type I error rates exceeding the .05 significance level for the additional 20 items with no DIF. This trend can be explained by how conditioning scores classify examinees at each score level to conduct DIF analysis. Traditional test score matching classifies heterogeneous people into the same category for DIF detection whereas attribute profile score matching resulted in more homogeneous groups when DIF was introduced through the DINA model by increasing the slip parameter and decreasing the guessing parameter for the focal group. Of the four DIF detection methods, MH-T had the greatest Type I error rates overall and was more prone to inflated Type I error rate problem under the influence of the sample size, amount of DIF, and types of DIF. MH-P was the most stable and conservative measure with lower Type I error rates for all test conditions.

Power Study

Power was defined as the rate of the correct identification of items simulated to have DIF. The percentages of detection of the five items that were simulated to have DIF out of the total counts across 25 replications were used as an empirical estimate of the power of the statistic. It should be noted that the interpretation of power was conditioned on the Type I error rates under the significance level because the power of hypothesis testing can be increased by the inflated Type I error rate. In the power study, the number of correctly identified DIF items was compared across various sample size, attribute pair-wise tetrachoric correlations, amounts of DIF, and types of DIF using MH statistic and SIBTEST with both the traditional test score matching and the attribute profile score matching.

MH Analysis with Two Matching Criteria

The results of the power study of MH-T and MH-P as a function of sample size, amount of DIF, attribute pair-wise tetrachoric correlations, and types of DIF are presented in Tables 9 and Table 10. Table 9 lists the power rates for MH statistics with test score (MH-T) and profile score matching (MH-P) when the amount of DIF in DIF-items was a .075 difference in both the slip and guessing parameters between the focal group and the reference group. Table 10 lists similar output only when the amount of DIF was increased to a .15. As sample size increased, power rates for both methods increased. When DIF amount was a .075 difference in the parameters, the average power rates increased from sample size 400 to 800 12.13% for MH-T and 14.63% for MH-P respectively. When the amount of DIF equaled a .15 difference in the parameters, MH-T had a similar average power increase (11.89%) as the MH-P (11.79%). On average, both methods had similar increases in power rate as the sample size increased (11.91% increase for MH-T and 13.26% increase for MH-P).

Table 9. Power Rates for the Five Items Averaged over 125 Counts (5x25) When the Amount of DIF =.075. (MH-P and MH-T)

		MH-profile		MH-testscore	
		Group size			
	Correlation	400	800	400	800
No DIF	ρ=0	.040	.040	.040	.040
	ρ=.5	.014	.064	.012	.056
	ρ=.8	.040	.048	.040	.064
Increasing slip	ρ=0	.112	.160	.080	.112
	ρ=.5	.104	.144	.104	.104
	ρ=.8	.168	.256	.128	.240
Increasing guessing	ρ=0	.320	.648	.232	.504
	ρ=.5	.344	.520	.296	.424
	ρ=.8	.176	.432	.160	.296
Increasing both slip and guessing	ρ=0	.280	.376	.200	.336
	ρ=.5	.088	.288	.080	.240
	ρ=.8	.048	.128	.056	.088
Increasing slip and decreasing guessing	ρ=0	.528	.840	.544	.768
	ρ=.5	.608	.848	.536	.720
	ρ=.8	.648	.920	.472	.808

Table 10. Power Rates for the Five Items Averaged over 125 Counts (5x25) When the Amount of DIF =.15. (MH-P and MH-T)

		MH-profile		MH-testscore	
		Group size			
	Correlation	400	800	400	800
No DIF	ρ=0	.040	.056	.056	.042
	ρ=.5	.048	.032	.048	.016
	ρ=.8	.072	.056	.032	.032
Increasing slip	ρ=0	.176	.272	.176	.264
	ρ=.5	.336	.544	.256	.408
	ρ=.8	.424	.728	.256	.488
Increasing guessing	ρ=0	.808	.952	.704	.912
	ρ=.5	.720	.968	.624	.896
	ρ=.8	.688	.88	.592	.816
Increasing both slip and guessing	ρ=0	.616	.768	.488	.648
	ρ=.5	.336	.584	.312	.504
	ρ=.8	.080	.272	.072	.288
Increasing slip and decreasing guessing	ρ=0	.992	1	.984	1
	ρ=.5	.992	.992	.984	1
	ρ=.8	.992	1	.976	1

The results show that as the amount of DIF increased, power rate also increased on average in the similar amount for both the MH-T (23.65% increase) and the MH-P (23.98% increase). Figure 13 shows the power rates for the different amounts of DIF and attribute pair-wise tetrachoric correlations averaged across sample sizes and types of DIF. MH-P yielded higher power rates than MH-T for both amounts of DIF. Higher amounts of DIF for both MH-T and MH-P yielded excellent power rates in combination with the two types of DIF(increasing the guessing parameter condition and the uniform DIF condition where p- values were increased for both the masters and nonmasters in the focal group) and a larger sample size (800/800 for the focal and reference groups).

Figure 13. MH Power Rates as a Function of the Amount of DIF and Attribute Pair-wise Tetrachoric Correlations

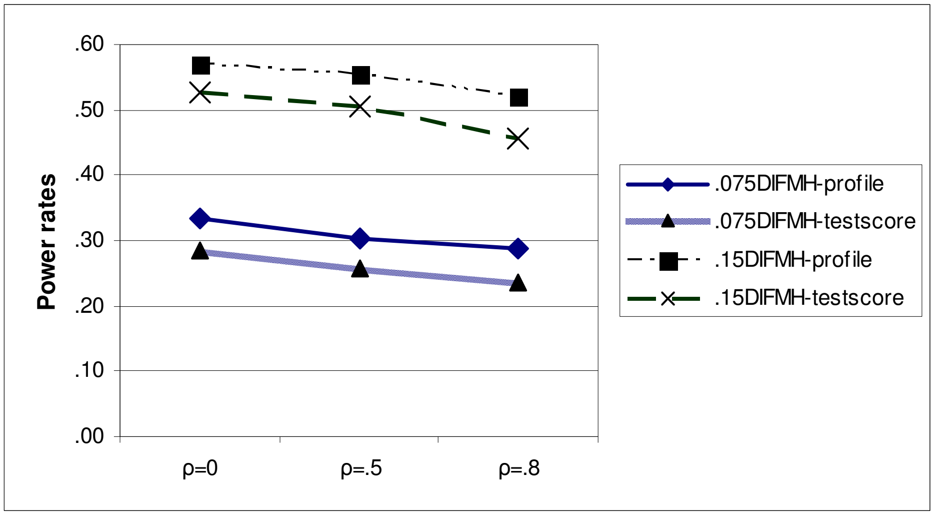
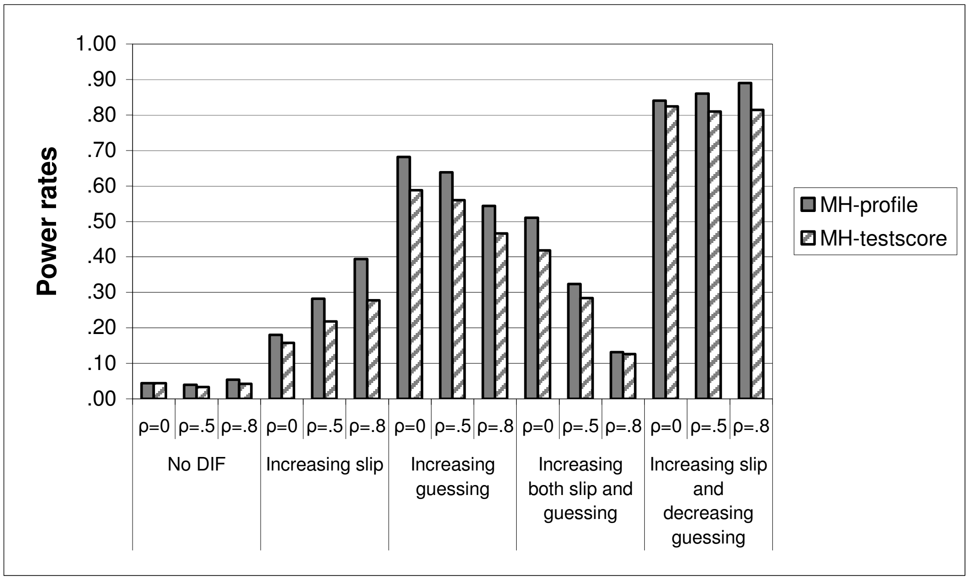


Figure 14 shows attribute pair-wise tetrachoric correlation had similar effect on the change of power rates for both the MH-P and MH-T.

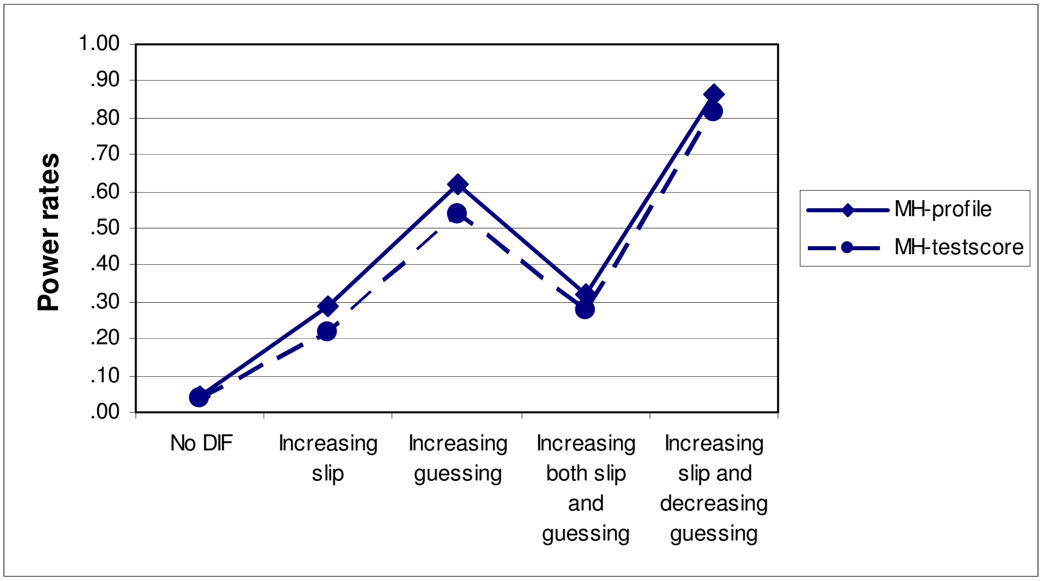
Figure 14. MH Power Rates Averaged across Sample Size and Amount of DIF



When only the slip parameter was changed, power rates for the two MH methods tended to increase as the examinee responses approximated unidimensionality. This increasing power pattern did not occur for the other types of DIF. When only the guessing parameter was increased and when nonuniform DIF were introduced by manipulating the slip and the guessing parameters in the same direction, the power rate decreased as the increasing attribute correlations approximated unidimensional item responses. In the last type of DIF case where uniform DIF was produced by keeping a lower probability of getting the correct response for the masters and nonmasters in the focal group compared to the reference group, the change in attribute correlations seemed to have minimal impact on the power rate for both matching criteria. There seemed to be an interaction effect between the attribute correlation and the types of DIF. Figure 14 demonstrated that when power rates were averaged across types of DIF and sample size, lower attribute association condition yielded the highest power, followed by the moderate correlation and the greatest correlation. It appears that both MH methods had higher average power in correctly identifying DIF items when the examinees demonstrate independent skills in solving the questions. Nevertheless, the change was negligible as the difference of power between attribute correlations did not exceed .05.

Figure 15 displays the power rates for two MH methods for the four types of DIF. Overall both MH methods appeared to have similar power in identifying the DIF items for four types of DIF. For all types of DIF, MH-P consistently had a higher power rate when compared to the MH-T.

Figure 15. MH Power Rates as a Function of Types of DIF



Types of DIF have a distinctive impact on the power of MH-T and MH-P. For both matching methods, when uniform DIF was introduced by decreasing the slip and increasing the guessing parameters where p-values were increased for the masters and nonmasters in the focal group, MH statistics with two matching criteria had the highest power rate for correctly identify the DIF items out of all five types of DIF. MH statistics with both matching criteria had low power rates of detecting DIF correctly (27.6% for MH-T and 34.2% for MH-P) when nonuniform DIF was present. However, the lowest power rates were observed when only the slip parameter was increased (21.8% for MH-T and 28.53% for MH-P). Increasing the guessing parameter alone yielded power rates of 53.80% for MH-T and 62.13% for MH-P.

Figure 16 through Figure 18 plot the power rates for both MH statistics calculated with two matching criteria across five types of DIF at each level of attribute correlation conditions. Power difference was expected to be observed between two MH methods when the correlation between attribute was low indicating a multidimensional item response structure. In that case, MH-P was expected to provide higher power rates when compared to MH-T. Similar power rates of MH-P and MH-T were expected to occur as the examinee responses approached unidimensionality with a high attribute correlation. However this pattern did not seem obvious in the simulation results. One reason might be related to the fact that attribute correlation had a minimal effect on the power, also interaction effect between types of DIF and attribute correlation influenced the magnitude of the power.

Figure 16. Power rates for MH-T and MH-P as a Function of Types of DIF when ρ=0

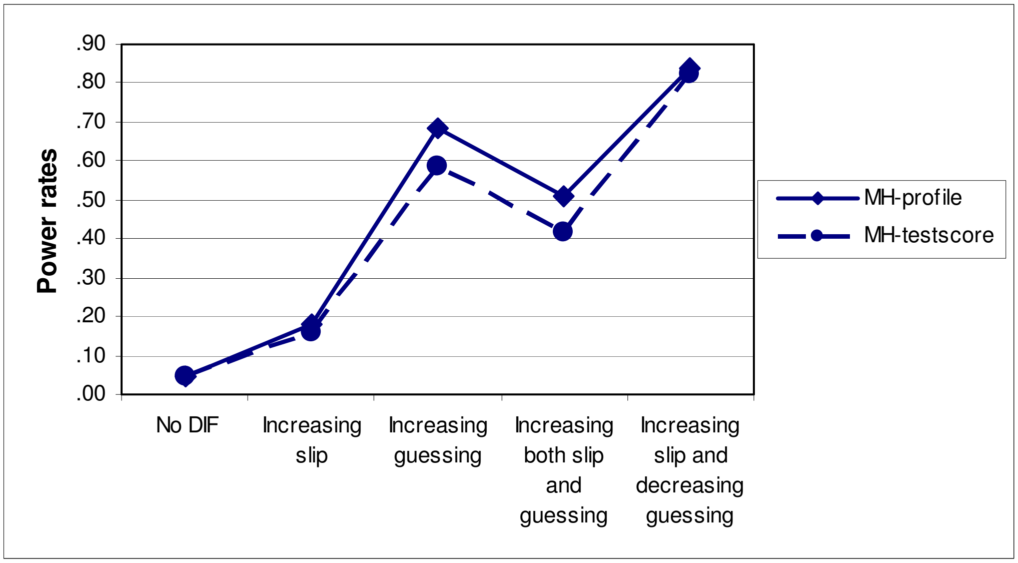


Figure 17. Power Rates for MH-T and MH-P as a Function of Types of DIF When ρ=.5

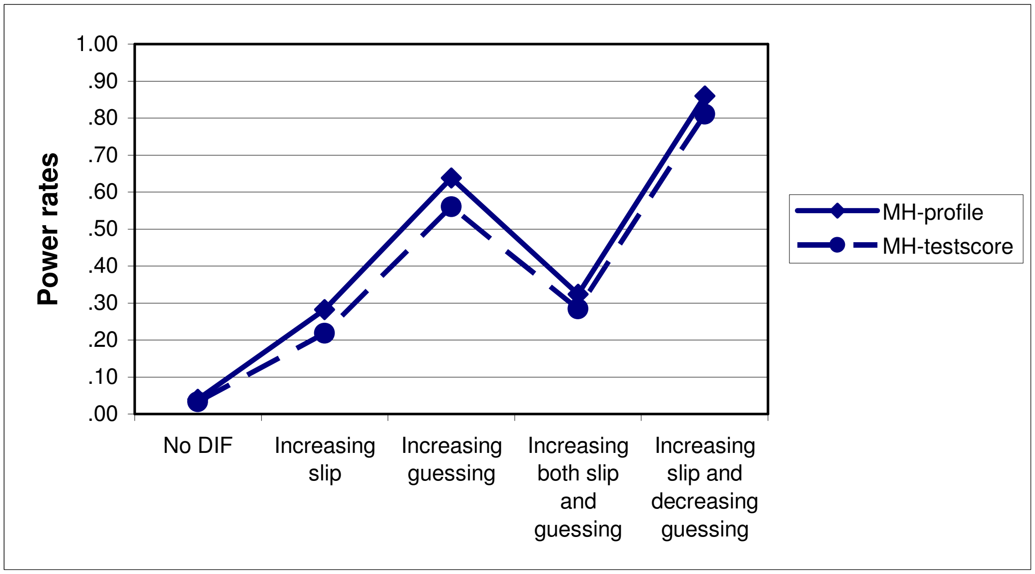
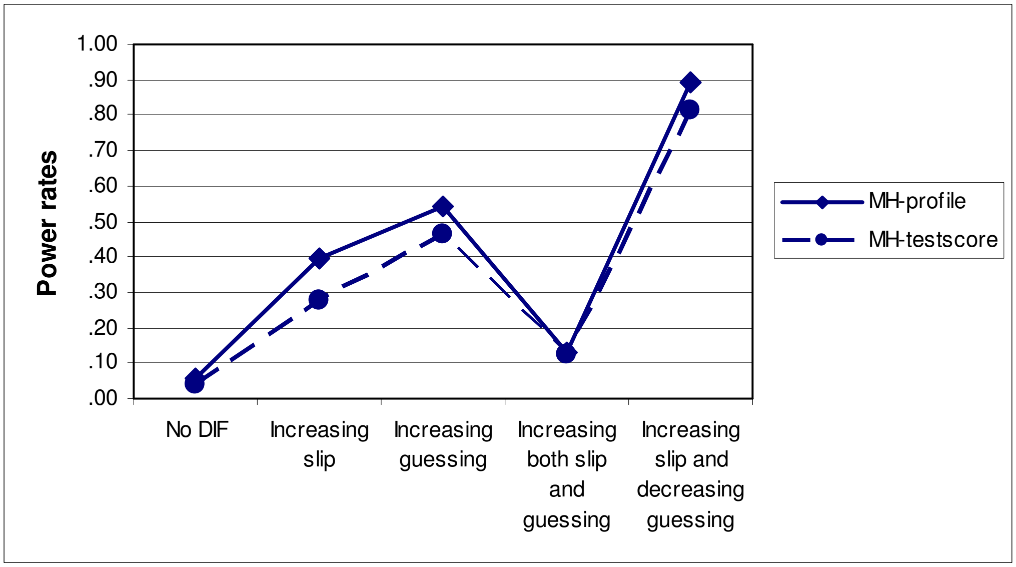


Figure 18. Power rates for MH-T and MH-P as a Function of Types of DIF When ρ=.8



SIBTEST with Two Matching Criteria

The results for power study of SIBTEST with true score (SIBTEST-T) and attribute profile score conditioning (SIBTEST-P) as a function of sample size, amount of DIF, attribute pair-wise tetrachoric correlations, and types of DIF are presented in Table 11 and 12. Table 11 lists the power study for SIBTEST-T and SIBTEST-P when the amount of DIF in DIF-items is a .075 difference in the slip and the guessing parameters between the focal group and the reference group. Table 12 lists similar output only when the amount of DIF was increased to a .15 difference.

Table 11. Power Rates for the Five Items Averaged over 125 Counts (5x25) When the Amount of DIF =.075. (SIBTEST-P and SIBTEST-T)

		SIB-profile		SIB-truescore	
		Group size			
	Correlation	400	800	400	800
No DIF	ρ=0	.056	.048	.072	.064
	ρ=.5	.020	.056	.032	.048
	ρ=.8	.048	.048	.040	.056
Increasing slip	ρ=0	.096	.144	.112	.120
	ρ=.5	.088	.136	.088	.144
	ρ=.8	.168	.264	.096	.264
Increasing guessing	ρ=0	.328	.680	.200	.328
	ρ=.5	.368	.544	.128	.232
	ρ=.8	.240	.480	.088	.120
Increasing both slip and guessing	ρ=0	.288	.376	.120	.192
	ρ=.5	.120	.296	.064	.104
	ρ=.8	.048	.144	.024	.056
Increasing slip and decreasing guessing	ρ=0	.560	.880	.296	.552
	ρ=.5	.600	.840	.312	.488
	ρ=.8	.696	.928	.296	.600

Table 12. Power Rates for the Five Items Averaged over 125 Counts (5x25) When the Amount of DIF = .15. (SIBTEST-P and SIBTEST-T)

		SIB-profile		SIB-truescore	
		Group size			
	Correlation	400	800	400	800
No DIF	ρ=0	.040	.064	.056	.064
	ρ=.5	.072	.048	.048	.032
	ρ=.8	.072	.056	.048	.064
Increasing slip	ρ=0	.176	.256	.136	.200
	ρ=.5	.344	.552	.328	.424
	ρ=.8	.344	.728	.336	.056
Increasing guessing	ρ=0	.824	.960	.616	.872
	ρ=.5	.800	.968	.448	.704
	ρ=.8	.760	.904	.360	.520
Increasing both slip and guessing	ρ=0	.584	.784	.400	.608
	ρ=.5	.384	.600	.128	.280
	ρ=.8	.144	.904	.024	.520
Increasing slip and decreasing guessing	ρ=0	.992	1	.896	.984
	ρ=.5	.984	.992	.864	1
	ρ=.8	.992	1	.864	1

When DIF amount was a .075 difference in the parameters, the average power rates increased from sample size 400 to 800 9.3% for SIBTEST-T and 14.27% for SIBTEST-P respectively. When the DIF amount was equal to a .15 difference, two SIBTEST methods had similar magnitude of power increases (11.84% for SIBTEST-T and 15.36% for SIBTEST-P). When the DIF amount was high, the power increased across sample size and was higher for both of the SIBTEST methods than when the DIF amount was low.

Table 13 lists the power rate difference between SIBTEST-P and SIBTEST-T as a function of amount of DIF and sample size. SIBTEST-P had greater power rates ranging from .117 to .166 than the SIBTEST-T in two sample size conditions.

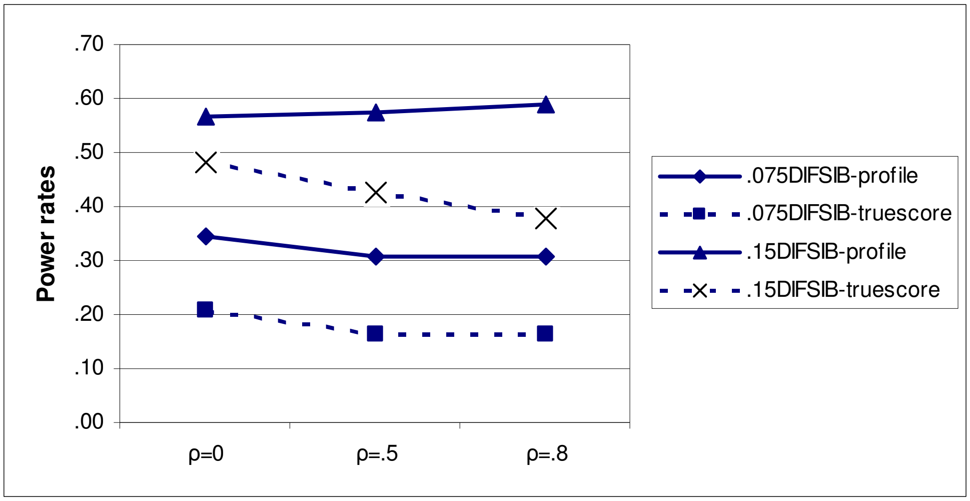
Table 13. The Means and Standard Deviations of SIBTEST Power Rates Difference as a Function of Sample Size and Amount of DIF

	(SIB-profile)-(SIB-truescore)	
Sample size	400	800
DIF amount=.075	.117 (.131)*	.166 (.159)
DIF amount=.15	.131 (.128)	.166 (.200)

*Standard deviations are given within the parentheses

Figure 19 demonstrates that as the amount of DIF increased, power rate increased for both SIBTEST-T (a 25.10% increase) and SIBTEST-P (a 25.80% increase).

Figure 19. SIBTEST Power Rates as a Function of the Amount of DIF and Attribute Pair-wise Tetrachoric Correlations



It was noted when the DIF amount was small, SIBTEST-T had probabilities below 50% in detecting the uniform DIF items correctly when averaged across sample size whereas the SIBTEST-P had some good power rates ranging form 72% to 81.2%. In uniform DIF case, both the masters and the nonmasters in the focal group had lower probability of getting correct responses than the reference group. When the DIF amount increased to a .15 difference, the results indicated that both SIBTEST methods had higher power in detecting DIF items in two DIF cases: the uniform DIF and changing only the guessing. Overall, the amount of DIF had similar effects on the power rate increase for both SIBTEST methods.

Figure 19 also shows, averaged across sample size and types of DIF, the attribute pair-wise tetrachoric correlations had a minimal impact on the power rates for SIBTEST with the two matching criteria. Especially, when the amount of DIF was large, SIBTEST-P yielded similar power rates across three attribute correlation conditions. In contrast, with the same amount of DIF, SIBTEST-T tended to yield lower power with a high attribute association which approximated unidimensional examinee responses compared to other levels of attribute associations.

The effect of the attribute correlation on the power rate was not consistent across types of DIF introduced through the DINA model for both SIBTEST matching methods. Table 14 lists the power rates of SIBTEST-T and SIBTEST-P as a function of sample size, types of DIF and attribute pair-wise tetrachoric correlations averaged over amounts of DIF.

Table 14. Power Rates as a Function of Sample Size, Types of DIF and Attribute Pair-wise Tetrachoric Correlations (SIBTEST-P and SIBTEST-T)

		SIB-profile		SIB-truescore	
		Group size			
	Correlation	400	800	400	800
No DIF	ρ=0	.048	.056	.064	.064
	ρ=.5	.046	.052	.040	.040
	ρ=.8	.060	.052	.044	.060
Increasing slip	ρ=0	.136	.200	.124	.160
	ρ=.5	.216	.344	.208	.284
	ρ=.8	.256	.496	.216	.160
Increasing guessing	ρ=0	.576	.820	.408	.600
	ρ=.5	.584	.756	.288	.468
	ρ=.8	.500	.692	.224	.320
Increasing both slip and guessing	ρ=0	.436	.580	.260	.400
	ρ=.5	.252	.448	.096	.192
	ρ=.8	.096	.524	.024	.288
Increasing slip and decreasing guessing	ρ=0	.776	.940	.596	.768
	ρ=.5	.792	.916	.588	.744
	ρ=.8	.844	.964	.580	.800

Table 14 shows that as examinees exhibited more unidimensional responses (generated by a higher attribute tetrachoric correlation), power rates for both SIBTEST tended to decrease when only the guessing parameter was changed for the focal group. In another case, where only the slip parameter was changed, the higher the attribute correlations were, the higher the power rates for both SIBTEST methods. When the p-values for both masters and nonmasters for the focal group were increased as in the uniform DIF case, the attribute correlation had little impact on power rates as the three correlations yielded similar power rates. In summary, the effect of attribute correlation was not consistent for all types of DIF.

Figures 20 – 22 displays the plots of the power rates for SIBTEST-P and SIBTEST-T across five types of DIF for each attribute correlation. Attribute tetrachoric correlations had little effect in minimizing the power rates difference between SIBTEST-P and SIBTEST-T. Across three attribute association levels, the power rates for SIBTEST-P was .25 higher than the counterpart SIBTEST-T in three types of DIF cases, changing the guessing parameter alone, nonuniform DIF and the uniform DIF. As examinee responses approximated unidimensional, SIBTEST-P started to show the same magnitude of power advantage over SIBTEST-T in changing only the slip parameter DIF case, indicating that SIBTEST-P appeared to have significant larger power (on average .25) than SIBTEST-T across all DIF types.

Figure 20. Power Rates for SIBTEST-T and SIBTEST-P as a Function of Types of DIF When ρ=0

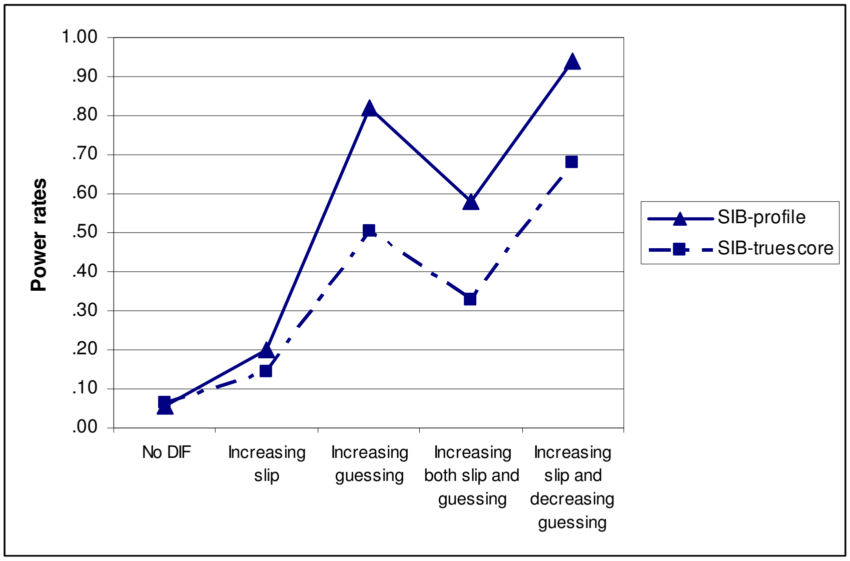


Figure 21. Power Rates for SIBTEST-T and SIBTEST-P as a Function of Types of DIF When ρ=.5

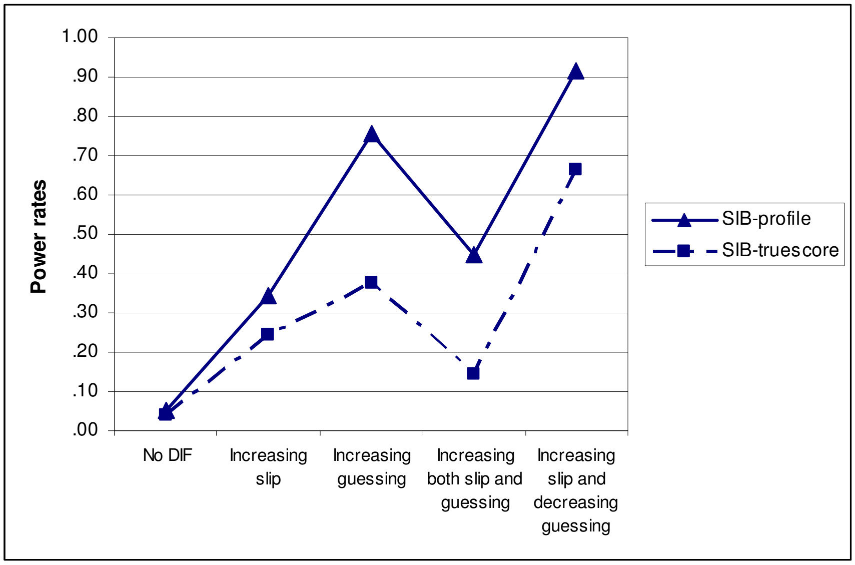
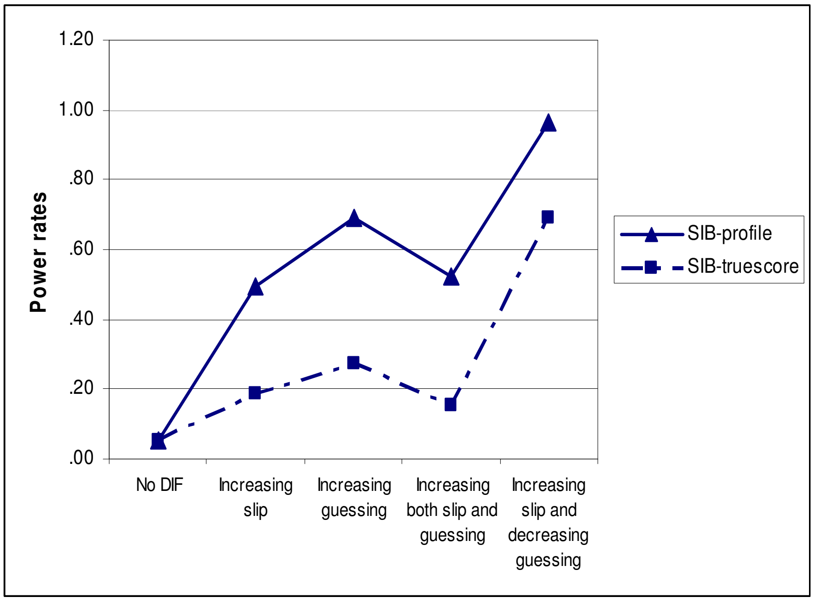


Figure 22. Power Rates for SIBTEST-T and SIBTEST-P as a Function of Types of DIF When ρ=.8



The types of DIF simulated through the DINA model had a great impact on the power rates for SIBTEST-T and SIBTEST-P. SIBTEST-P had a higher power rate than SIBTEST-P in four types of DIF conditions. Overall, SIBTEST-T had poor power rates in detecting DIF items correctly in all other types of DIF conditions except when uniform DIF was present. For example, SIBTEST-T had moderate power rates ranging from 74.4% to 80% in the uniform DIF condition for 800/800 sample size condition. The power rates for SIBTEST-P peaked for the uniform DIF condition ranging from 91.6% to 96.4% for the same sample size. When only the guessing parameter was increased, SIBTEST-P yielded moderate power rates ranging from 69.2% to 82% when sample size was 800 for across all attribute correlation conditions. Changing only the slip parameter did not result in satisfying power rates for two SIBTEST methods. In addition, both

SIBTEST methods did poorly in identifying nonuniform DIF where differences between masters and nonmasters in the focal group might be cancelled out. Despite this fact, in nonuniform DIF type, SIBTEST-P still yielded twice the power rates as SIBTEST-T.

Comparison between MH and SIBTEST

The results of the power study comparison for MH-P and SIBTEST-P are listed in Table 15 as a function of the sample size, attribute pair-wise tetrachoric correlations and types of DIF conditions.

Table 15. Power Rates Averaged across Two Amounts of DIF (SIBTEST-P and MH-P)

		SIB-profile		MH-profile	
		Group size			
	Correlation	400	800	400	800
No DIF	ρ=0	.048	.056	.040	.048
	ρ=.5	.046	.052	.031	.048
	ρ=.8	.060	.052	.056	.052
Increasing slip	ρ=0	.136	.200	.144	.216
	ρ=.5	.216	.344	.220	.344
	ρ=.8	.256	.496	.296	.492
Increasing guessing	ρ=0	.576	.820	.564	.800
	ρ=.5	.584	.756	.532	.744
	ρ=.8	.500	.692	.432	.656
Increasing both slip and guessing	ρ=0	.436	.580	.448	.572
	ρ=.5	.252	.448	.212	.436
	ρ=.8	.096	.524	.064	.200
Increasing slip and decreasing guessing	ρ=0	.776	.940	.760	.920
	ρ=.5	.792	.916	.800	.920
	ρ=.8	.844	.964	.820	.960

In general, the number of correctly identified DIF items was greater for SIBTEST-P (54.77%) compared to when MH-P (52.30%) was used over the 60 test conditions. When uniform DIF was introduced into the data, both the MH-P and

SIBTEST-P tended to yield moderate to excellent power rates for the larger sample size, a trend that was also observed for the SIBTEST-T and MH-T. The correlations between attributes had minimal impact on the power rates for the two methods. However, there was an interaction effect between attribute correlations and types of DIF. When only the slip parameter was manipulated and when uniform DIF was introduced, power rates for both methods tended to increase as the correlation was increased. An opposite trend existed when only the guessing parameter was changed and nonuniform DIF case was present. Under those two types of DIF conditions, both methods resulted in decreasing power rates when examinee responses approximated unidimensionality with higher tetrachoric attribute correlations. When the attribute correlations were held constant as illustrated in Figure 23 - 25, as examinee responses approximated multidimensionality indicated by a low attribute association, both methods provided similar power rates in detecting DIF items. When examinee responses approximated undimensionality, a discrepancy (.15 difference) in power between SIBTEST-P and MH-P occurred in the nonuniform DIF case. However, in the nonuniform DIF case, both methods had really low power. When uniform DIF was present, regardless of the attribute association, the two methods performed equally well with excellent power rates.

Figure 23. Power Rates for SIBTEST-P and MH-P as a Function of Types of DIF When ρ=0

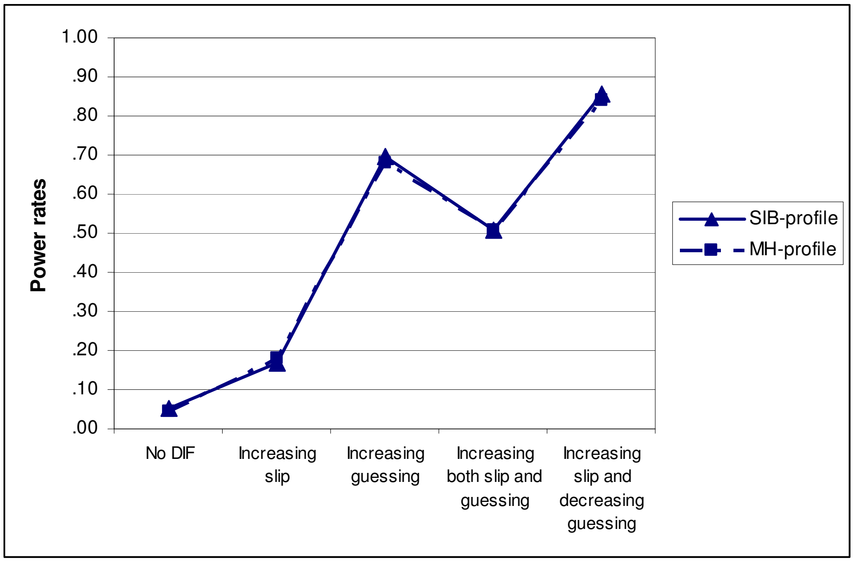


Figure 24. Power Rates for SIBTEST-P and MH-P as a Function of Types of DIF When ρ=.5

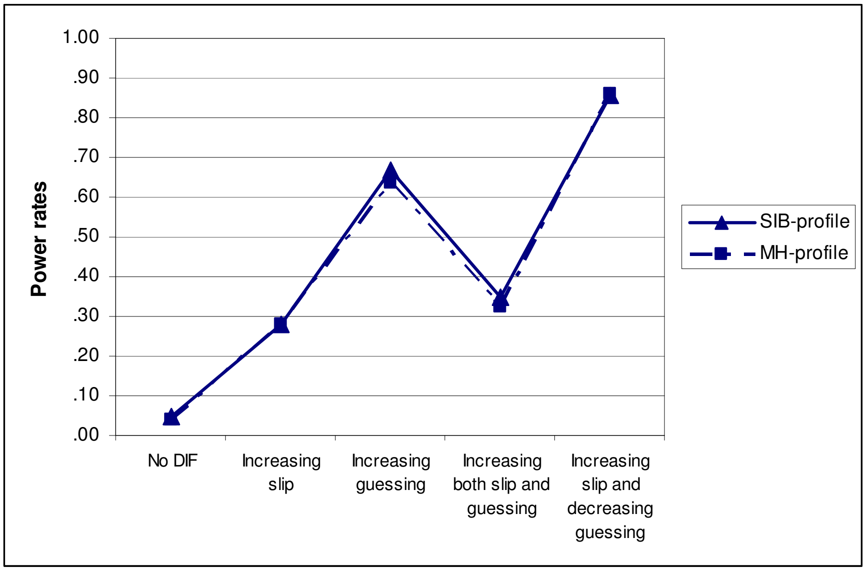
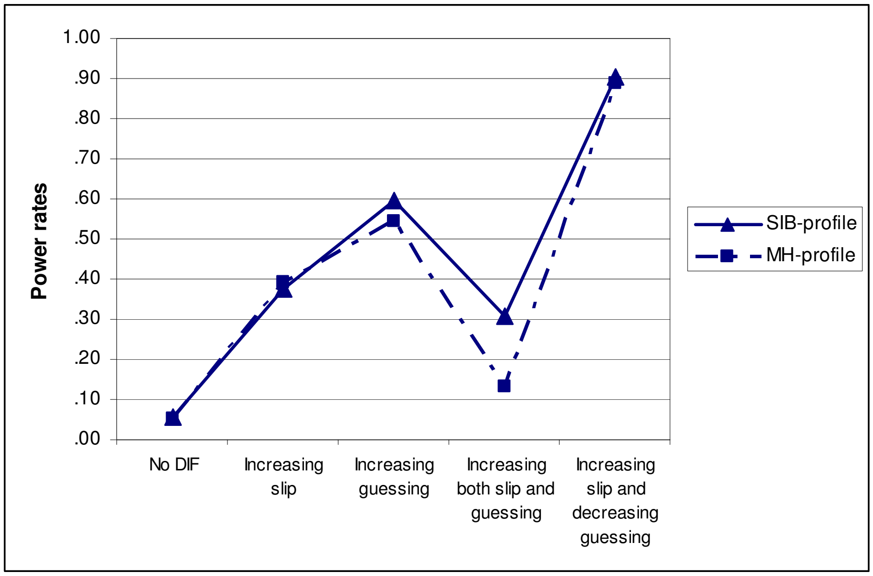


Figure 25. Power Rates for SIBTEST-P and MH-P as a Function of Types of DIF When ρ=.8



In summary, despite the evidence of exhibiting low power in some DIF types, the MH and SIBTEST with profile score matching provided consistent power rates under the studied conditions. Neither of the MH-P and SIBTEST-P appeared to be sensitive in detecting nonuniform DIF items. For the two particular DIF types that included changing only the guessing parameter and the uniform DIF cases, both methods yielded moderate to good power rates in detecting DIF items.

The results for power study comparison for MH-T and SIBTEST-T are listed in Table 16 as a function of sample size, attribute correlation and types of DIF conditions.

Table 16. Power Rates Averaged across Two Amounts of DIF (SIBTEST-T and MH-T)

		SIB-truescore		MH-testscore	
		Group size			
	Correlation	400	800	400	800
No DIF	ρ=0	.064	.064	.048	.041
	ρ=.5	.040	.040	.030	.036
	ρ=.8	.044	.060	.036	.048
Increasing slip	ρ=0	.124	.160	.128	.188
	ρ=.5	.208	.284	.180	.256
	ρ=.8	.216	.160	.192	.364
Increasing guessing	ρ=0	.408	.600	.468	.708
	ρ=.5	.288	.468	.460	.660
	ρ=.8	.224	.320	.376	.556
Increasing both slip and guessing	ρ=0	.260	.400	.344	.492
	ρ=.5	.096	.192	.196	.372
	ρ=.8	.024	.288	.064	.188
Increasing slip and decreasing guessing	ρ=0	.596	.768	.764	.884
	ρ=.5	.588	.744	.760	.860
	ρ=.8	.580	.800	.724	.904

Moderate to excellent power rates from SIBTEST-T and MH-T were observed for uniform DIF conditions with a larger sample size. The correlations between attributes had minimal impact on the power rates for the two methods. When the attribute correlations were held constant, as illustrated in Figures 25 to 27, as the examinee responses approximated multidimensionality (indicated by a low attribute association), the advantage of MH-T over SIBTEST-T in power rate was the same for the three DIF types, namely, changing only the guessing parameter, nonuniform DIF and uniform DIF. When examinee responses approximated unidimensionality (indicated by a high attribute association), the power difference became larger for the two types of DIF: changing only the slip parameter and changing only the guessing parameter.

Figure 26. Power Rates for SIBTEST-T and MH-T as a Function of Types of DIF When ρ=0

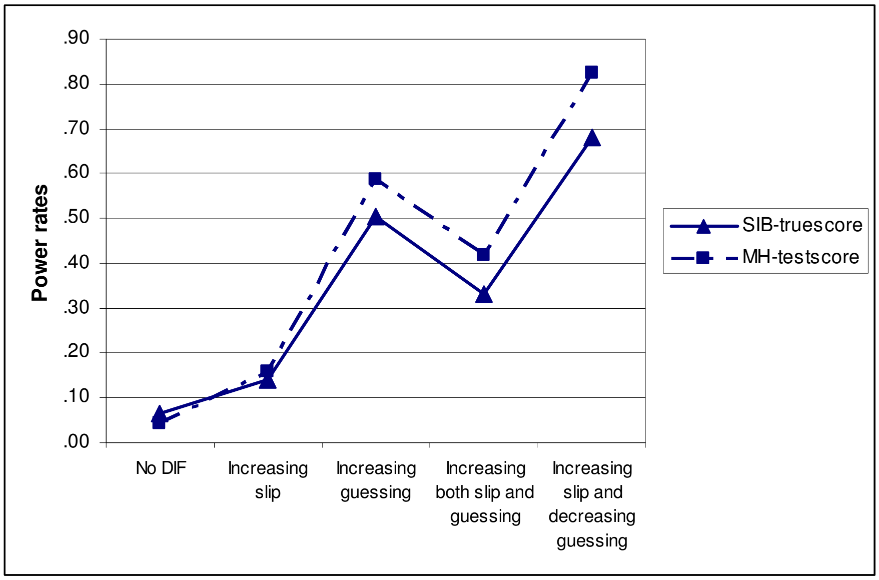


Figure 27. Power Rates for SIBTEST-T and MH-T as a Function of Types of DIF When ρ=.5

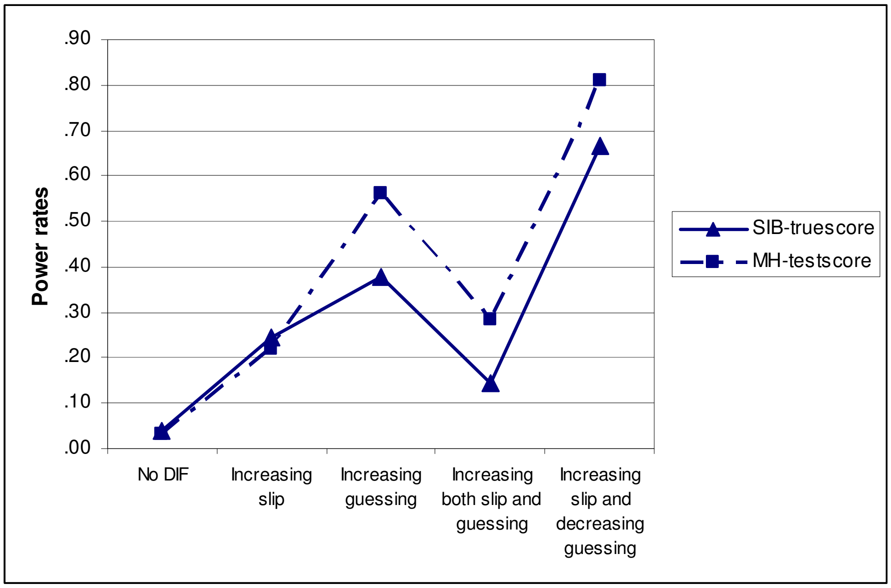
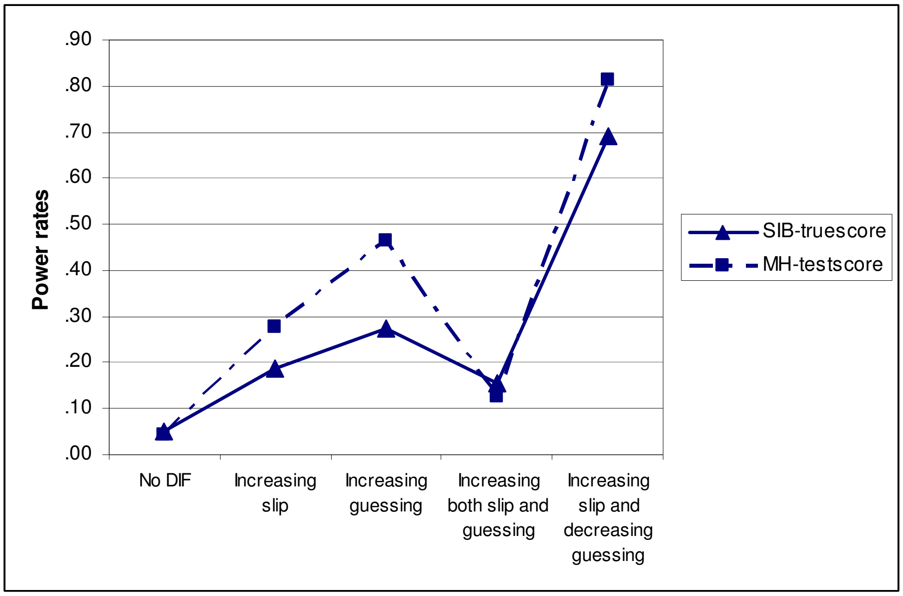


Figure 28. Power Rates for SIBTEST-T and MH-T as a Function of Types of DIF When ρ=.8



In summary, MH-T appeared to result in higher power rates than SIBTEST-T. However considering MH-T also yielded higher Type I error rates compared to the Type I error rates from SIBTEST-T under specific test conditions, MH-T did not necessarily appear to perform any better than SIBTEST-T in detecting DIF items.

CHAPTER V

REAL DATA APPLICATION

In this section, an example of applying DIF detection procedures using attribute profile score matching was compared to DIF detection procedures using traditional test score matching with real data. Currently, few tests have been developed with cognitive diagnosis objectives in mind. Attribute profile score matching is only valid and practical when test items are written with a cognitive diagnostic model as the correct model. However, real data applications can provide the researcher insight for comparing the performance of profile score matching and traditional test score or adjusted true score matching when conditioning on examinees' assessed primary ability to detect DIF.

MH statistic and SIBTEST statistic based on total test score matching (MH-T and SIBTEST-T) and attribute profile score matching (MH-P and SIBTEST-P) were used to examine differential item functioning (DIF) on a dataset from the 1999 Trend of International Math and Science Study (TIMSS). Using students' self-reported gender information, statistical DIF analyses were conducted for the gender groups. Females were designated as the reference group and the males as the focal group.

A total of 1132 examinees at 8th grade took the 14-item booklet developed to measure math and science abilities. After deleting the subjects that failed to provide gender information and item responses, a total of 1104 examinees were retained for the final analysis. Among the examinees, there were 559 female and 545 males. The items for the TIMSS data can be found in Table 17, the first six items were science items, and the last eight items were math items. Because the TIMSS was not developed with a cognitive diagnosis purpose in mind, a Q-matrix was not supplied. A Q-matrix developed by Templin (2004) for TIMSS data was adopted which had four attributes, each taken from the TIMSS data item content listing. The attribute descriptions for the item content Q-matrix can be found in Table 18. The entries of the item content Q-matrix can be found in Table 19. The Q-matrix limited each item only measuring one attribute or one skill, which was commonly referred to as “simple structure”.

Table 17. TIMSS Test Items

1. The picture shows the three main layers of the Earth. Where is it the hottest?
2. Most of the chemical energy released when gasoline burns in a car engine is not used to move the car, but is changed into?
3. Which object listed in the table has the greatest density?
4. Immediately before and after running a 50 meter race, your pulse and breathing rates are taken. What changes would you expect to find?
5. The diagram below shows a mountain. The prevailing wind direction and average air temperatures at different elevations on both sides of the mountain are indicated. Which feature is probably located at the base of the mountain at location X?
6. The walls of a building are to be painted to reflect a much light as possible. What color should they be painted?
7. According to the information in the graph, during which two-month period does the greatest increase in coat sales occur?
8. If there are 300 calories in 100 g of a certain food, how many calories are there in a 30g portion of this food?
9. Which picture shows that $2/5$ is equivalent to $4/10$?
10. Which of these is the smallest number?
11. Which of these cubes could be made by folding the figure above?
12. *n* is a number. When *n* is multiplied by 7, and 6 is then added, the result is 41. Which of these equations represents this relation?
13. A club has 85 members, and there are 14 more girls than boys. How many boys and how many girls are members of the club? (free response item)
14. A sheet of paper is 0.012 cm thick. Of the following, which would be the height of a stack of 400 sheets of this paper?

Table 18. TIMSS Item Content Q-matrix Attribute Descriptions

A1. Earth/life science.
A2. Physics.
A3. Data and Fractions.
A4. Geometry and Algebra.

Table 19. TIMSS Item Content Q-matrix

Item	A1	A2	A3	A4
1	1	0	0	0
2	0	1	0	0
3	0	1	0	0
4	1	0	0	0
5	1	0	0	0
6	0	1	0	0
7	0	0	1	0
8	0	0	1	0
9	0	0	1	0
10	0	0	1	0
11	0	0	0	1
12	0	0	0	1
13	0	0	0	1
14	0	0	0	1

A summary of descriptive statistics for the test and subgroups is reported in Table 20.

Table 20. TIMSS Descriptive Statistic

Examinee	n	Mean	SD	Reliability
Female	559	8.62	2.46	.60
Male	545	8.81	2.71	.68
Total	1104	8.71	2.59	.64

The TIMSS data item parameters and examinee parameter were calibrated using DINA MCMC software developed for this study. MCMC with Metropolis-Hasting within the Gibbs Sampler iterations were set at 20000 runs with the first 10000 runs used as the burn-in. The long iteration chains yielded more stable and accurate item and examinee parameters. The slip and guessing parameter estimates are listed in Table 21.

Table 21. TIMSS Item Parameters

Item	s	$1-s$	g
1	.04	.96	.73
2	.31	.69	.55
3	.66	.34	.12
4	.02	.98	.56
5	.47	.53	.26
6	.05	.95	.75
7	.14	.86	.54
8	.19	.81	.57
9	.25	.75	.45
10	.29	.71	.14
11	.14	.86	.40
12	.03	.97	.61
13	.44	.56	.10
14	.18	.82	.36

Unlike the ranges for the slip and guessing parameters used in the simulation study, the slip parameter estimates for this dataset ranged from .033 to .660 and the guessing parameters ranged from .115 to .748. These were still reasonable estimates except that higher guessing parameters indicated that nonmasters appeared to get higher probabilities of a correct response compared to the simulation guessing parameters. The $1-s_j$ column represented the p-values for the masters in both subgroups. If small values were found in this column, it could indicate that some skills were not defined in the Q-matrix. Easier items normally have low slip parameters but high guessing parameters. Difficult items have high slip parameters and low guessing parameters. Average difficulty items have both low slip and guessing parameters. Items in this TIMSS data appeared to be easy or average difficulty items except that Item 3 was a relatively hard item with high slip parameter and low guessing parameter.

There were a total of nine attribute mastery patterns estimated by the DINA MCMC program. Table 22 lists the frequencies of the observed attribute mastery patterns. Table 23 lists the frequency of the observed total test scores. These score categories were later used as examinees' ability conditioning for the DIF analysis on the gender groups. The frequency distribution of the raw score was negatively distributed indicating this was an easy test.

Table 22. The Frequency of the Attribute Mastery Pattern

Attribute Mastery Pattern				Frequency
0	0	0	0	90
0	0	1	0	3
0	0	1	1	3
0	1	1	1	1
1	0	0	0	417
1	0	0	1	1
1	0	1	0	81
1	0	1	1	71
1	1	1	1	437

Table 23. The Frequency of Total Test Score

Test Score	Frequency
0	2
1	3
2	5
3	15
4	31
5	68
6	104
7	132
8	146
9	160
10	139
11	132
12	90
13	58
14	19

It was noted that with TIMSS data there were fewer profile score categories than the observed total test score categories. Because the same cognitive process may still yield correct or incorrect responses and different cognitive process could yield the same correct answer, classifications as a result of latent cognitive constructs were different from those of the observed test scores. There was no direct link as to the equivalence of total test score with the profile score categories, which indicated the two matching methods represented different perspectives in describing examinee's ability. MH-T, MH-P, SIBTEST-T and SIBTEST-P analyses were conducted to summarize the similarities and differences in identifying DIF items on gender groups. MH detection procedure used Chi-square statistics and SIBTEST used two-tailed Z-statistics. DIF statistics (MH statistics and SIBTEST) identified similar DIF items with two matching criteria.

Table 24 shows the same three items (1, 9 and 12) out of 14 items had p-values less than the significance level of .05 based on four DIF methods. All four DIF detection methods demonstrated that Item 1 favored the male group whereas Item 9 and Item 12 favored the female group. In terms of the magnitude of the effect size in measuring DIF according to ETS guidelines on $\hat{\beta}_{uni}$, all three items were identified with large DIF using SIBTEST-P, while SIBTEST-T identified Item 9 and Item 12 as items with large DIF and Item 1 with moderate DIF. Following the ETS delta metric, an effect size measure for MH statistic, MH statistics based on both matching criteria identified Item 1 and Item 12 as items with large DIF and Item 9 as an item with moderate DIF.

Table 24. DIF Statistics for MH and SIBTEST with Two Matching Criteria

Item No.	MH-profile		MH-testscore		SIB-profile	SIB-truescore
	χ^2	Δ_{MH}	χ^2	Δ_{MH}	$\hat{\beta}_{uni}$	$\hat{\beta}_{uni}$
1	12.294**	2.303	13.275**	2.405	-3.643**	-.059**
9	11.787**	-1.043	12.528**	-1.16	3.405**	.106**
12	17.218**	-1.596	16.698**	-1.773	4.137**	.096**
2	6.348*	0.754	5.808*	0.765	-2.61**	-0.065
4	7.29**	-3.402	1.021	-0.715	3.007**	0.02

Note: * denotes p<.05 and ** denotes p<.01

Table 25. Selected DIF Items

1. The picture shows the three main layers of the Earth. Where is it the hottest?
9. Which picture shows that $2/5$ is equivalent to $4/10$?
12. *n* is a number. When *n* is multiplied by 7, and 6 is then added, the result is 41. Which of these equations represents this relation?
2. Most of the chemical energy released when gasoline burns in a car engine is not used to move the car, but is changed into?
4. Immediately before and after running a 50 meter race, your pulse and breathing rates are taken. What changes would you expect to find?

Table 25 lists the item content for selected DIF items. An examination on Table 25 with item context for the first DIF items revealed that all three items were associated with picture reading or equation reading skills. It was possible that a secondary ability difference between female and male groups, which was not measured by the test, existed in the examinee responses when DIF occurred. Neither the total test score nor the attribute profile score can account for this ability dimension because the auxiliary skill wasn't counted towards total test score and wasn't listed in the Q-matrix for attribute mastery pattern estimation.

In addition, SIBTEST-P and MH-P identified additional items with DIF (Item 2 and Item 4). Item 4 was not classified as DIF item by MH-T and SIBTEST-T. Both MH-T and MH-P identified Item 2 as DIF-item with p-values less than the significance level. SIBTEST-T also yielded a p-value on Item 2 just slightly above the .05 significance level. Item 2 measured the physics content, but examinees were likely to mistaken it for a chemistry or a mechanical question because the terms such as chemistry energy and gasoline were mentioned in the item. The SIBTEST-P was a standardized mean difference method based on profile score matching, that didn't resemble the regular

SIBTEST-T. Compared with SIBTEST-T, MH-T and MH-P, SIBTEST-P detected larger degree of DIF among the four approaches. When this specific dataset had fewer profile score categories, SIBTEST-T appeared to be the least sensitive measure by identifying fewer DIF items. MH-P yielded smaller effect size measures than the MH-T given significant p-values, indicating MH-P did not tend to overestimate the presence of DIF compared to MH-T.

The purpose of this real data application was to introduce a new matching criterion based attribute profile score for investigating DIF with cognitive diagnostic framework. MH statistics and SIBTEST statistics with two matching criteria were compared. A Q-matrix was developed for the TIMSS data to illustrate the underlying skills measured by the items (Templin, 2004). It was hypothesized that the profile score matching would be a valid matching criterion if it could better classify people into more homogeneous skill groups. However the real data application on profile score matching with its limited profile score categories did not account for the examinees' ability as much as was expected, hence DIF procedures with profile score matching identified more items with DIF. Another factor related to the performance of the profile score matching method was the construction of the Q-matrix where skills were specified before constructing the test. For the purpose of this study, a Q-matrix was developed after the test was constructed and represented only one example of the many ways of interpreting the skills measured in the test. The construction of the Q-matrix had a direct effect on the estimation of the profile score categories.

Despite the limitations, profile score matching for most cases provided similar DIF results as the traditional total test score matching and the adjusted true score matching in MH statistic and SIBTEST. Under constrained situations incorporating the Q-matrix construction, DIF detection procedures with profile score matching could serve as the an additional tool to examinee subgroup differences with the same skill mastery pattern.

CHAPTER VI

CONCLUSIONS

This dissertation has demonstrated how DIF detection procedures can be used in conjunction with the cognitive diagnosis framework to detect potential difference between subgroups after accounting for their measured primary ability (i.e., the attribute profile scores). The small step in integrating cognitive theories with statistical DIF analyses indicated that it was possible to manipulate DIF through differences in item parameters from the DINA model across the levels of grouping variable and to use examinee attribute profile scores as a matching criterion to detect the created DIF. The simulation study examined the performance of two DIF detection procedures (MH and SIBTEST) with three matching criteria (test score, true score and profile score matching). Type I error rate and power rate were used to assess the impact of different test conditions on DIF procedures by systematically varying the sample size of the subgroups, the attribute pair-wise tetrachoric correlations (i.e., to vary the degree of dimensionality), types of DIF introduced by varying the parameters in the DINA model and the amount of difference between item parameters.

The simulation results of this study appeared to confirm that profile score matching for MH statistics and SIBTEST could accurately classify people into the same ability group and were robust in maintaining the Type I error rate where factors such as sample size, DIF amount, attribute associations and different types of DIF were varied.

When the MH statistic conditioned upon profile matching it seemed to be more conservative when compared to SIBTEST that also conditioned upon profile matching, resulting in better Type I error control. Both methods with profile score matching were consistent in identifying DIF items with equal power for the uniform DIF case. Neither SIBTEST-P nor MH-P was sensitive in detecting nonuniform DIF. However, SIBTEST-P was different from the previous SIBTEST research because it did not adopt the mechanism of traditional SIBTEST in adjusting the examinees' true score based on the reliability of the observed scores. The SIBTEST-P was reduced to a standardization mean difference method with profile score matching.

Sample size, amount of DIF and the type of DIF each affected the Type I error rate and the power. For example, a large increase in power occurred for MH-P when the sample size and amount of DIF were increased. The effect of attribute correlation on the DIF detection statistics appeared to be negligible when other test conditions were held constant. Nevertheless, when examinee responses approximated unidimensionality, SIBTEST-P and SIBTEST-T literally had the lowest yet similar Type I error rates among other attribute association conditions where uniform DIF type was present. Neither procedure overestimated the presence of DIF for a specific attribute association level or type of DIF.

When the examinee responses were approximately multidimensional, as indicated by a low attribute correlation, SIBTEST-T, SIBTEST-P and MH-P had the same power in detecting DIF-items for each type DIF. All three methods performed equally well in matching examinees. Both SIBTEST-P and MH-P used examinees' profile (e.g., latent multidimensional ability) as the conditioning variable. Similarly, SIBTEST-T using adjusted true score that also represented the examinees' ability in a multidimensional composite. It was expected that profile attribute score matching would lose its superiority to test score matching in detecting DIF items when the examinee responses approached unidimensionality because the total test score was sufficient to account for examinees' primary ability. Only MH-P and MH-T had similar power for uniform DIF in the high attribute correlation condition. However, the results of this study did not suggest a definite pattern between SIBTEST-P and SIBTEST-T. In contrast, the Type I error rate for MH-T appeared to be influenced more by the type of DIF, introduced by varying the parameters in the DINA model, in combination with a larger sample size and a greater amount of DIF.

When changing only the guessing parameter and introducing uniform DIF, Type I error rates for the MH statistic based on either test score or profile score matching were inflated in contrast with other types of DIF. The power increase was also observed in the same two DIF cases. The same pattern occurred for SIBTEST based on two matching criteria, however, the amount of increase in Type I error rate was not as much. The reason why specific two types of DIF had more effect on Type I error rate and power than the others was related to the fact that in both situations significant DIF for both the masters and nonmasters was created in the subgroups. In contrast, when the slip parameter was increased, only the p-value for the masters in the focal group was decreased. The result of Type I error rate and power rate indicated this type of DIF was not different from the No-DIF situation. The finding might be due to the specification of

Q-matrix, which restrained the proportion of examinees that could be classified as the masters. Because of a low volume of masters in the focal group, any difference that was created for the masters in the two subgroups became negligible.

The difference and similarity between two matching criteria was investigated through a simulation study and a real data analysis. DIF detection procedures using attribute profile score matching captured and reflected the item parameter differences introduced using the DINA model. The sensitivity of the DIF methods appeared to rely on the estimation of the examinee attribute profile scores that was also influenced by the Q-matrix construction. Any other skills that were not specified in the Q-matrix would result in DIF when conditioning examinees on their attribute profile scores. It appeared that the traditional MH statistic and SIBTEST still functioned sufficiently well in the cognitive diagnostic framework despite the evidence that profile score matching was superior to the traditional matching criteria in terms of keeping Type I error around the significance level and yielding good power. However, profile scores were not as straightforward as obtaining the total test scores, that involved the assumption of the cognitive structure and cognitive diagnostic model selections. Practitioners should identify the purpose of the test and select DIF matching criterion accordingly.

The popularity of the cognitive diagnostic models and their application in large scale assessment makes it necessary to bridge the gap between applying traditional DIF procedures and investigating DIF from a cognitive diagnostic perspective. More ideas about future studies arise from this exploratory work of defining DIF using a cognitive diagnostic model.

First, only the DINA model was used to simulate DIF conditions, which did not differentiate between examinees missing one attribute and examinees mastering none. In the DINA model framework, examinees were either categorized into the masters or the non-master groups. As was found in the study, varying the difference for the nonmasters had a more significant impact on the detection of DIF than only varying the difference for the masters. The classification of masters and nonmasters would influence the correctness of the DIF procedures based on profile score matching. Comparisons between other cognitive diagnostic models that recognize the difference between examinees mastering some attributes and missing all attributes need to be pursued. Second, it was assumed that both the focal group and the reference group had equal number of examinees, which in reality may not always be the case. The factor of sample size needs to be further investigated. Third, profile score categories were influenced by the number of attributes in the Q-matrix. When attributes increase for the test, the categories will increase exponentially. The current study used a single number of five attributes and a fixed test length of 25 items. There were more total test score categories than attribute profile score categories. Future study should investigate unequal sample sizes, different number of attributes and tests of different lengths. As cell counts have a direct effect on the significance of the MH chi-square test, a large number of categories with fewer cell counts may subsequently decrease the power of the statistic. Also in real data applications, examinees only exhibited certain patterns of attribute mastery that could be far fewer than the possible total number of test score categories. In that case, the constraint from Q-matrix played the role in defining on what mastery level DIF detection procedures can be performed. In this study it appeared that the real data Q-matrix configuration actually constrained the number of observed possible profile patterns. It is suggested that the profile score matching obtained from cognitive diagnostic model not be the only matching criterion for DIF detection procedures before more elaborate research are conducted on test developed for cognitive diagnostic purpose.

Last, future study should relate cognitive diagnostic models with multidimensional IRT models with a joint effort in understanding the dimensionality association between attributes. The current simulation does not emulate the multidimensionality through examinee attribute response pattern as efficiently as other research where multidimensional IRT models may be adopted. It is hoped that after the measured skills are accounted for in the cognitive diagnostic models, the degree of DIF will be decreased and the cause of DIF can be explained through examining the Q-matrix specification where the attribute and the item relationship are defined.

To practitioners this thesis has demonstrated DIF can be created using a cognitive diagnostic model from a skill mastery perspective and can be detected by conditioning on estimated attribute profile scores. If a cognitive diagnostic model appears to be the “correct” model then subsequent DIF analyses using profile matching appear to be a viable alternative. For skill assessment, in addition to performing nonparametric DIF detection procedures, it provides valuable insights to investigate whether there is a large difference in the item parameters, such as the slip and the guessing parameters, between the subgroups. When differences in the item parameters are observed for both masters and nonmasters it appears that DIF is likely to occur. After a careful examination on the item context and cross check with the Q-matrix, practitioners might be able to identify whether DIF is caused by additional dimensions not identified in the Q-matrix or if groups of examinees are employing strategies that are not accounted for by their resulting score profile.

REFERENCES

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.

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

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord &M. R. Novick, *Statistical theories of mental test scores* (pp. 395–479). Reading MA: Addison-Wesley.

Bolt, D. M. (2002). A Monte Carlo Comparison of Parametric and Nonparametric Polytomous DIF Detection Methods. *Applied Measurement in Education*. Vol. 15(2), 113–141.

Clauser, B. E., Mazor, K. & Hambleton, R. K. (1993). The effect of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, Vol. 6 Issue 4, 269-279.

Clauser, B. E., Nungester, R. J., Mazor, K. & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement*, Summer96, Vol. 33 Issue 2, 202-215.

Clauser, B., Nungester, R.,& Swaminathan (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement*, Vol. 33, No. 4, 453-464.

Cohen, J. (1992). A power primer. *Psychological Bulletin*,112, 155-159.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P.D. Nichols, S. F. Chipman, & R.L. Brennan. *Cognitively diagnostic assessment* (pp.361-389). Hillsdale, NJ; Erlbaum.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*. V69, 3, 333-353.

Douglas, J., Roussos, L. A. & Stout, W. F. (1996). Item-Bundle DIF Hypothesis Testing: Identifying Suspect Bundles and Assessing Their Differential Functioning. *Journal of Educational Measurement*, V33 n4 p465-84.

Embretson, S. (1997). Multicomponent response models. In W.J. van der Linden & R.K. Hambleton (Eds,). *Handbook of modern item response theory* (pp.305-321). New York: Springer-Verlag.

Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of test* (pp. 359-384). Hillsdale, NJ: Erlbaum.

Hartz, S. (2002). Skills diagnosis: Theory and practice. User Manual for Arpeggio software. Princeton, NJ: ETS.

Henson, R., & Douglas, J. (2004). Test Construction for Cognitive Diagnosis. *Applied Psychological Measurement, 29 (4)*, 262-277.

Henson, R., & Templin, J. (2006). Implications of Q-matrix misspecification in cognitive diagnosis models. Manuscript under review.

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

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.

Hambleton, R. K., & Jones, R. W. (1994). Comparison of empirical and judgmental procedures for detecting differential item functioning. *Educational Research Quarterly*, 18, 21-36.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14(4), 329-349.

Junker, B. W., & Sijtsma, K. (2001). Cognitive Assessment Models with few Assumptions, and Connections with Nonparametric item Response Theory. *Applied Psychological Measurement, 25(3)*, 258-272.

Junker, B. W. (2001). On the interplay between nonparametric and parametric IRT, with some thoughts about the future. In A. Boomsma, M. A. J. Van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 274–276). New York: Springer-Verlag.

Kim, S. H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, 8, 291-312.

Kubiak, A., O'Neill, K., & Payton, C. (1992). The effects of using educational background variables in DIF analysis. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

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.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64, 187-212.

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.

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: the effects of matching on unidimensional subtest scores. *Journal of Educational Measurement*, 32, 131-144.

Mislevy, R. J. (1997). Probability-based Inference in Cognitive Diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43-71). Hillsdale, NJ: Erlbaum.

No Child Left Behind Act of 2001 (NCLB) Public Law 107-110.

Nandakumar, R., & Roussos, L. (2001). CATSIB: a modified SIBTEST procedure to detect differential item functioning in computerized adaptive tests. Law School Admission Council computerized testing report, 97-11.

Oshima, T. C., & Miller, M. D. (1990). Multidimensionality and IRT-based item invairaince indexes: The effect of between group variation in trait correlation. *Journal of Educational Measurement*, 27, 273-283.

Oshima, T. C., & Miller, M. D. (1990). Multidimensionality and Item Bias in item response theory. *Applied Psychological Measurement* 16, 237-248.

Oshima, T. C., McGinty, D., & Flowers, C. P. (1994). Differential item functioning for a test with a cutoff score: Use of limited closed-interval measures. *Applied Measurement in Education*, 7, 195-209.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146–178.

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

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement* v 21 n1 p25-36.

Roussos, L. A. & Stout, W. F. (1996). Simulation Studies of the Effects of Small Sample Size and Studied Item Parameters in SIBTEST and Mantel-Haenszel Type I Error Performance.; *Journal of Educational Measurement*, v33 n2 p215-30 Sum.

Rudas, T., & Zwick, R. (1997). Estimating the importance of differential item functioning. *Journal of Educational and Behavior Statistics*, Vol 22, 1, 31-45.

Ryan, K. E. (1991). The Performance of the Mantel-Haenszel Procedure across Samples and Matching Criteria. *Journal of Educational Measurement*, v28 n4 325-37 Win.

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.

Sinharay, S., Almond, R., & Yan, D. (2004). Model checking for models with discrete proficiency variables in educational assessment (ETS RR-04-04). Princeton, NJ: ETS.

Spiegelhalter, D. J., Best, N. G., Gilks, W. R., & Inskip, H. (1996). Hepatitis B: A case study in MCMC methods. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 21–43). London: Chapman and Hall.

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*, 12, 55-73.

Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & Safto, M. (Eds.). *Monitoring skills and knowledge acquisition* (pp.453-488). Hillsdale, NJ; Erlbaum.

Tatsuoka, K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P.D. Nichols, S. F. Chipman, & R.L. Brennan. *Cognitively diagnostic assessment* (pp.327-359). Hillsdale, NJ; Erlbaum.

Templin, J. (2004). Generalized Linear Mixed Proficiency Models for Cognitive Diagnosis. Manuscript under Review.

Yamamoto (1989). HYBRID model of IRT and latent class models. Research Report. RR-89-41. Princeton, NJ: Educational Testing Service.

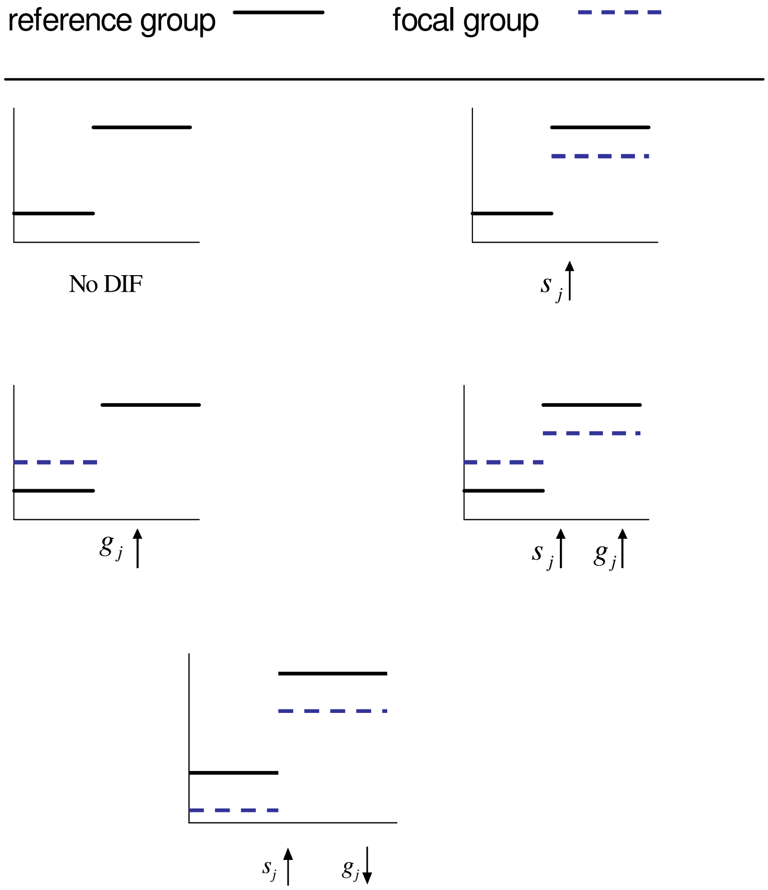
Zenisky, A.L., Hambleton, R.K., & Robin, F. (2003). Detection of Differential Item Functioning in Large-Scale State Assessments: A Study Evaluating a Two-Stage Approach l. *Educational and Psychological Measurement*, Vol. 63 No. 1, 51-64.

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

Zwick, R, & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, 26, 55-66.

APPENDIX

GRAPHICAL ILLUSTRATION


reference group
focal group
No DIF
$s_j \uparrow$
$g_j \uparrow$
$s_j \uparrow$ $g_j \uparrow$
$s_j \uparrow$ $g_j \downarrow$