This is to certify that the
dissertation entitled

USING THE MULTIVARIATE MULTILEVEL
LOGISTICREGRESSION MODEL TO DETECT DIF: A
COMPARISON WITH HGLM ANDLOGISTIC REGRESSION
DIF DETECTION METHODS

presented by

Tianshu Pan

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Measurement and Quantitative Methods

Mark Reckase

Major Professor's Signature

12-10-08

Date

MSU is an Affirmative Action/Equal Opportunity Employer

USING THE MULTIVARIATE MULTILEVEL LOGISTIC REGRESSION MODEL TO DETECT DIF: A COMPARISON WITH HGLM AND LOGISTIC REGRESSION DIF DETECTION METHODS

By

Tianshu Pan

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods

2008

ABSTRACT

USING THE MULTIVARIATE MULTILEVEL LOGISTIC REGRESSION MODEL TO DETECT DIF: A COMPARISON WITH HGLM AND LOGISTIC REGRESSION DIF DETECTION METHODS

By

Tianshu Pan

This study presents the Multivariate Multilevel Logistic Regression (MMLR) models to detect Differential Item Functioning (DIF), which are likely to detect DIF when the responses of an examinee are not locally independent. The study also compares the uses of the three MMLR models, three modified versions of Kamata's Hierarchical Generalized Linear Model (HGLM) and the standard logistic regression model as DIF detection methods. The comparison between these statistical procedures for DIF detection will be made using Michigan Educational Assessment Program reading test and simulated data. The simulation study evaluates their performances in the detection of uniform DIF. Simulated data are generated by the 3-parameter logistic Item Response Theory models, varying conditions of different sample size (400, 700, and 1000 examinees), test length (20, 40 and 60 items), the difference of parameter *b* (0.25, 0.50, and 0.75) and the ability distributions with different means and variances for the reference and focal groups. These test conditions are crossed completely and replicated 500 times. In these analyses, total score and IRT ability estimate are respectively used as the matching variable. The results show that MMLR can be used for DIF detection. It is also found that the heterogeneous variances of the two groups influence power and Type I error rates of these methods, and the HGLM DIF models are unsuitable to identify DIF.

This dissertation is dedicated

to

my grandmother

Tao Cheng (陶诚).

ACKNOWLEDGEMENT

I appreciate very much many persons because this dissertation could not be completed without their help, assistance and encouragement. I am deeply thankful to Dr. Mark Reckase, the chair of my dissertation committee for his direction and patience throughout my doctoral studies. I wish to express my gratitude to my other committee members, Dr. Tenko Raykov, Dr. Connie Page, and Dr. Yuehua Cui for their comments. In addition, I am also thankful for the help of Dr. Yeow Meng Thum and Dr. Kimberly Maier.

Finally, I am greatly indebted to my family. I thank my father, Pan Konggen, and mother, Sun Baozhen. They have always supported me and encouraged me to pursue a Ph.D. degree in the U.S.A. I also thank my wife, Chen Yumin, who did most of the chores which enabled me to complete my doctoral study and dissertation, and my daughter, Pan Yuqi, whose loveliness made me forget any annoyance and hardships with my study and work. I give my most special thanks and long yearning to my grandmother, Tao Cheng, who took care of me from when I was born until when I got married. Unfortunately she did not have a chance to see me study abroad and complete the dissertation. I hope she is proud of me in Heaven and happily knows that I have finished my Ph.D. studies.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1
Introduction

In this chapter, the background and the importance of this study will be described first. Then, related literature is also reviewed. The chapter will make clear what is new in the study and state the purposes of the study.

1.1 Background

Sometimes, examinees in different demographic groups who have the same levels of ability have different probabilities of answering a particular item correctly. The difference is defined as differential item functioning (DIF). Differential item functioning (DIF) refers to differences in the functioning of an item among groups after the groups have been matched with respect to the ability or attribute that the item purportedly measures (Dorans & Holland, 1993) In the item response theory (IRT) framework, DIF means that the test item has a different item response function for one examinee group than for others (Lord, 1980) and it can be defined as a difference in the conditional probabilities that persons of the same ability answer an item correctly in two or more groups (Hidalgo & López-Pina, 2004). DIF is viewed as a necessary but not a sufficient condition for item bias (Clauser & Mazor, 1998).

In the case of two groups, they are identified as reference and focal groups. The reference group is composed of the majority or advantage group while the focal group is composed of the minority or disadvantage group and this group is considered the subject of DIF analysis. Mellenberg (1982) classified DIF as uniform and nonuniform DIF. In the framework of IRT models, uniform DIF occurs when the item characteristic curves (ICC) for the two groups differ only in the difficulty parameter and the relative advantage for the reference or focal group is uniform across the score scale. Nonuniform DIF is present when their ICCs are different as a result of the disparate differences of the discrimination parameters and/or pseudo-guessing parameters (Clauser & Mazor, 1998). Statistically, uniform DIF exists when there is no interaction between ability level and group membership. Nonuniform DIF exists when there is interaction between ability level and group membership. This study will compare the performance of three types of methods for detecting uniform DIF.

In the 1980s and the beginning of the 1990s, many DIF detection procedures were developed to identify DIF, such as Mantel-Haenszel (MH) (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), standardized difference (Dorans & Kulik, 1983; 1986), SIBTEST (Shealy & Stout, 1993) and so on. But these mentioned procedures can analyze only one item or small numbers of related items at a time (Swanson et al, 2002). Confirmatory Factor Analysis was used to identify all DIF in a test at a time (Muthén & Lehman 1985). But a simulation study shows the method has extremely low power to find DIF (Finch & French, 2007). Since 1990's, some types of the Generalized Linear Mixed Models (GLMM) were applied to identify DIF, for example, the Hierarchical Generalized Linear Model (HGLM) (Kamata, 1998; 2001), the Hierarchical Logistic Regression Model (HLRM) (Swanson et al, 2002) and the Logistic Mixed Model (LMM) (Van den Noortgate & Boek, 2005). The three DIF approaches are able to analyze all items in one computer run. Kamata's HGLM approach has different equations or mathematical forms shown by different authors. Kamata's HGLM and Van den Noortgate's LMM approaches both have random person ability. But random person ability makes them unsuitable to detect DIF when the examinees of the two groups have the different expected ability or proficiency. This study will analyze Kamata's model and try modifying it in order that it can be used when the two groups have different ability means. Swanson's HLRM method is not able to include the variations of the proficiency of examinees from different groups, e.g. classes or schools. In addition, for all of the methods mentioned here, it is impossible to take into account the probable dependence between the binary responses of the same examinee.

In order to address these disadvantages of these methods, a Multivariate Multilevel Logistic Regression (MMLR) Model will be introduced to detect DIF in this study.

1.2 The Statement of Purpose

The multivariate logistic regression model in this study is not the one which is usually used for the analysis of multinomial or ordinal data, but is an extension of the Multivariate Multilevel Linear Model. The model is presented by Griffiths et al (2004), Mcleod (2001), and Yang et al (2000). It is known as the Multivariate Multilevel Logistic Regression Model (MMLR) since it has at least two levels.

The main purposes of this study are to introduce the MMLR model to identify uniform DIF, set up a MMLR DIF detection procedure, and compare the performances of the MMLR, HGLM and LR DIF detection procedures identifying uniform DIF by a simulation study and their application to the real test data. These DIF detection approaches are compared since all of them use a logit transformation. Additionally, LR may be one of standard DIF detection methods as it is presented in the "Test Fairness" chapter (Camilli, 2006) of the book, *Educational Measurement*, sponsored jointly by National Council on Measurement in Education and American Council on Education. It is expected that MMLR is acceptable as a DIF detection procedure if it performs as well as LR when it is applied to detect DIF, otherwise it is not acceptable.

Second, as noted, there are confusing equations and a potential problem in Kamata's HGLM DIF procedure, which will be shown in the study. So, the secondary purpose of the study is to modify Kamata's HGLM DIF procedure to extend its applied conditions.

Chapter 2

DIF Detection Methods

This chapter is the related literature review about DIF detection methods. It gives more details about the DIF approaches and their disadvantages mentioned in Chapter 1. First, the chapter shows how to classify the current DIF detection methods, and then gives the details of some methods, and their disadvantages.

2.1 Classifications of DIF Detection Methods

Dozens of DIF procedures have been presented in the literature. They have been grouped under two major types by Camilli (2006). One is the use of IRT models and the other relies on analyzing the observed scores. The former includes Lord's difference in IRT parameters (Lord, 1980), Raju's signed and unsigned area indexes (Raju, 1988; 1990), the likelihood ratio tests (Thissen et al, 1988) and so on. The latter includes the Mantel-Haenszel, logistic regression, standardized Difference, SIBTEST, HGLM, and MMLR. The last method will be presented in this study. "While IRT methods provide useful results when the item models fit the data and a sufficient sample size exists for obtaining accurate estimates of IRT parameters, observed-score methods are frequently used with smaller sample sizes" than the IRT methods (Camilli, 2006: p. 236).

Among the IRT DIF methods, the IRT models for the reference and focal groups are estimated separately first, and then the differences between the item response functions of the two groups are calculated and tested for each item, e.g. Raju's indexes; or the differences of the parameters are computed and tested, e.g. Lord's approach. But in the likelihood ratio test, the likelihood ratio of the two IRT models is calculated for each item, i.e. the models with and without DIF, and has a large-sample chi-square distribution (Camilli, 2006). Since the proposed method in the study belongs to the observed-score methods, more details of the IRT methods are not shown in the dissertation.

2.2 Some Standard observed-score DIF Detection Methods

Here the methods that appear in the chapter by Camilli (2006) are labeled as standard methods. In this type of methods, scored item responses from the focal group are compared with the ones from the reference group in order to identify items that function differently in the two groups, using one or more additional covariates to control for individual differences on the construct to be measured. The covariate is called the matching variable or conditioning variable. Usually total raw score is used as a matching variable since it is easy and convenient to get the score.

2.2.1 The Mantel-Haenszel Procedure

The Mantel-Haenszel procedure was introduced to identify DIF by Holland and Thayer (1988). The MH DIF statistic is computed by matching examinees in each group on total test score and then forming *2x2xA* contingency tables for each item, where *A* is number of the score levels on the matching variable which is usually total test score. At each score level *S*, a 2-by-2 contingency table is created for each item,

	Correct	Incorrect	Total
Reference Group	C_{RS}	I_{RS}	n_{RS}
Focal group	C_{FS}	I_{FS}	n_{FS}
Total	C_{TS}	I_{TS}	n_{TS}

where C_{RS} stands for the number of reference group examinees at score level S who answer the item correctly. The other variables in the table have similar definitions. Then the effect size measure of DIF is obtained by

$$\hat{\alpha}_{MH} = (\sum_{S} C_{RS} I_{FS} / n_{TS}) / (\sum_{S} C_{FS} I_{RS} / n_{TS}) .$$

The statistic is typically converted to the log-odds scale, i.e. $\hat{\delta}_{MH} = \log(\hat{\alpha}_{MH})$. At Education Testing Service, it is put on the delta scale with the transformation:

$$MH\ D - DIF = \Delta_{MH} = -2.35 \hat{\delta}_{MH} .$$

Zieky (1993) divided the DIF magnitude into three categories according to the magnitude of Δ_{MH}.

2.2.2 The Logistic Regression DIF Detection Method

Swaminathan and Rogers (1990) first introduced logistic regression (LR) to detect DIF. They also showed that the Mantel-Haebszel (MH) procedure can be considered as being based on a LR model when the ability variable is discrete and no interaction term between the group variable and ability is specified. The logistic regression procedure employs the item response as the dependent variable, with a group membership variable, the abilities of examinees and the interaction between them as independent variables. The standard logistic regression model is expressed as

$$\ln(\frac{p_j}{1-p_j}) = \beta_0 + \beta_1 W_j + \beta_2 G_j + \beta_3 W_j G_j \qquad \textbf{(1)}$$

where p_j is the probability of examinee j's answering an item correctly, Wj is the matching variable, and could be the ability estimate or total score of examinee j, and G_j is the group membership of examinee *j*. The regression coefficients in the above equation can be estimated using maximum likelihood and can be tested for significance. If the item is unbiased, only β_0 and β_1 should be significantly different from zero. If β_2 is nonzero and β_3 equals zero, an item shows uniform DIF. If the interaction parameter β_3 is nonzero, the item has nonuniform DIF whether the other coefficients are equal to zero or not. Generally, total raw scores are used to indicate the proficiency of examinees. When the differentiating factors are assumed to function in the different patterns for examines with the same characteristics in different units, e.g. classes or schools, the standard logistic regression DIF model can be extended to a multilevel logistic regression model (e.g. van den Bergh et al, 1995).

2.2.3 The Standardized Difference Method

The standardized difference approach was introduced to analyze DIF by Dorans and Kulik(1983,1986). First, they calculate:

$$\Delta p_S = p_{RS} - p_{FS} = C_{RS} / N_{RS} - C_{FS} / N_{FS},$$

using the similar notation as the used for the contingency table of MH. Then, after these individual differences are summarizing across the levels of matching variables by applying some standardized weighting function to the differences, a standardized p-difference (STD P-DIF) can obtain be obtained by

$$STD\ P - DIF = (\sum_S w_S \Delta p_S) / \sum_S w_S .$$

The weight w_S can be defined in several ways. When the numbers of examinees at level S in the focal groups, n_{FS}, are applied, the standard error of *STD P-DIF* was given as follows by Dorans and Holland (1993: p.50).

$$SE(STDP-DIF) = \sqrt{P_F(1-P_F)/N_F + \sum_S n_{FS}^2 p_{RS}(1-p_{RS})/(n_{RS}N_F^2)}$$

Where P_F is the total correct proportion observed in the focal group, and N_F is the number of persons in the focal group.

2.2.4 The SIBTEST Procedure

The Simultaneous Item Bias Test (SIBTEST) was proposed by Shealy and Stout (1993). In the SIBTEST approach, test items are assigned to two subsets, the matching subtest and the suspect subtest. The suspect subtest could be a single item or bundles of items. In the former case, SIBTEST will analyze *n* items of a test individually and successively. On each trial, the i^{th} item is the object of study and the other $(n-1)$ items compose the matching subtest. The basic index for SIBTEST, *B*, is the mean of the group difference in subtest scores across the focal group ability distribution and is given by

$$B = \sum_S p_S(\overline{Y}_{RS}^* - \overline{Y}_{FS}^*).$$

where p_S is the proportion of focal group examinees among all focal group examinees and $\overline{Y}_{RS}^*, \overline{Y}_{FS}^*$ are the average item scores for the reference and focal group at the S^{th} level of the matching variable. An asymptotically normal test is provided by the ratio of *B* and its standard error, namely,

$$SIB = \frac{B}{Std\ Err(B)}$$

These procedures mentioned here are typically performed for each item individually or for small numbers of related items (Swanson et al, 2002). Typically, MH, LR and standardized difference just can deal with items individually. SIBTEST can be used to detect DIF either in single items or in bundles of items. But when it analyze bundles of items, these items are viewed as the whole and then it should be called Differential Bundles Functioning (Douglas et al, 1996) instead of DIF. Testing all items in one analysis not only intends to reduce the number of operations, but also make the procedure give us the results after analyzing the responses of persons to all items comprehensively. The following methods can address the problem.

2.3 The DIF Detection Method Based on Factor Analysis

Factor-analytic DIF approaches are able to analyze all items of a test in one run. The method can be traced to the 1970's (Humphreys & Taber, 1973). Later, Muthén and Lehman (1985) applied confirmatory factor analysis to look for the DIF items, which is based on the method of Muthén and Christoffersson (1981). Typically a first model is fit, with all factor parameters freely estimated across groups. This analysis provides several fit statistics, for instance, a chi-square goodness of fit test statistic. Then, a second constrained model is fit, with the hypothesis that all parameters (e.g. factor loadings) are equal across all groups, and a chi-square statistic is computed. The difference of the two chi-square statistics is calculated and compared to the critical value of the chi-square distribution with appropriate degrees of freedom in order to test for parameter invariance.

If the hypothesis is rejected in terms of the statistical significant result, further tests are implemented to single out items that contribute heavily to the rejection (Finch & French, 2007).

One disadvantage of the method is that no provision is made for the possibility of guessing (Muthén & Lehman, 1985). It may be allowed since the presence of guessing also influences the performances of other DIF detection procedures, such as SIBTEST and LR (e.g. Finch & French, 2007), according to some simulation studies. However, the largest disadvantage is that it has extremely low power (about 0.06) for DIF detection (Finch & French, 2007), even for some other measurement invariance detection (French & Finch, 2006) when the observed variables are dichotomous.

2.4 The DIF Detection Methods Based on GLMM

Since 1990s, some methods were developed to analyze all items in one run using the Generalized Linear Mixed Models, e.g. the Hierarchical Generalized Linear Model (HGLM), the Hierarchical Logistic Regression Model (HLRM) and the Logistic Mixed Model (LMM).

Kamata (1998, 2001) extended HGLM with 2 or 3 levels to set up a generalized Rasch Model and a HGLM DIF model. Subsequently, Luppescu (2002) and Kim (2003) respectively applied the method to identify uniform and nonuniform DIF, and Shen (1999) used a 3-level HGLM approach to detect DIF. Swanson et al (2002) developed a HLRM DIF detection approach. The details of Kamata's model will be given later and Swanson's model is shown as follows:

$$
\begin{aligned}
&\text{logit}[\text{Prob}(\text{Y}_{\text{ji}} = 1)] = b_{0i} + b_{1i}(\mathit{proficiency})_j + b_{2i}G_j \\
&b_{0i} = \gamma_{00} + u_{0i} \\
&b_{1i} = \gamma_{10} + u_{1i} \\
&b_{2i} = \gamma_{20} + \gamma_{21}I_1 + \ldots + \gamma_{2k}I_k + u_{2i}
\end{aligned}
$$

where Y_{ji} is the examinee *j*'s score on item *j*; $(\text{proficiency})_j$ is an index of proficiency on a common scale for all examinees; G_j is a dummy variable for the group membership; b_{0i} reflects (the log odds of) item difficulty in the reference group; b_{1i} reflects item discrimination, constrained (in this model) to be equal in reference and focal groups; and b_{2i} reflects the deviation of item difficulty in the focal group from the reference group; *I* is the dummy variable to indicate each item.

There are the following obvious distinctions between the two models:

- The persons are within an item in Swanson's model while the items are nested within a person in Kamata's method.
- Swanson's model has three random effects but Kamata's has only one.
- Swanson's model is based on a 2PL IRT model (Swanson et al, 2002) while Kamata's is Rasch-styled (Kamata, 1998; 2001).
- Swanson's model has a matching variable but Kamata's does not (the reason will be given later).

Van den Noortgate and Boek (2005) employed the logistic mixed model to detect uniform DIF, treating the person ability, the item effects and the interactions between items and groups as random. So, their model may be more appropriate to identify DIF for multiple groups than two groups of examinees.

Although Kamata's HGLM and Van den Noortgate's LMM approaches can deal with all items in an analysis but their uses may be limited as there is actually no matching variable in his model (the details will be given later in the dissertation). Swanson's HLRM procedure has a matching variable, but it is difficult to extend the model to 3-level model by including the variations for students from different classes or schools since the item level is the second level in the model. Additionally, all of these methods are implemented under the IRT assumption — local independence, which means that the responses of the same examinee are locally independent. It is impossible for them to deal with the local dependence, which possibly appears when a test consists of several testlets or measures multiple constructs.

A Multivariate Multilevel Logistic Regression Model can be applied to detect DIF and solve these problems. Although MMLR also belongs to the Generalized Linear Mixed Models family, MMLR could have a complex covariance matrix between the dichotomous responses within a cluster, which will be discussed in the next chapter. Then it is possible for MMLR to include the correlations between the responses in the model.

Chapter 3

Multivariate Multilevel Models

This chapter will introduce the Multivariate Multilevel Linear Model, generalize the model to the Multivariate Multilevel Logistic Regression Model, and apply it to detect DIF.

3.1 The Development of Multivariate Multilevel Models

Multilevel or Hierarchical Linear Models (HLM) appeared in the 1980's when contextual analysis and mixed effects models came together (Snijders & Bosker, 1999, *pp.* 1-2). Then the multilevel structure was used to construct multivariate multilevel models. There are a couple of different types of multivariate multilevel models. Goldstein & McDonald (1988), Goldstein (1995), Longford (1993), and Snijders & Bosker (1999) showed how to extend the hierarchical structure to do multivariate regression analysis for continuous dependent variables. Raudenbush and Bryk (2002: p. 450-54) and Hox (2002: p.158-161) presented the same hierarchical/multilevel multivariate linear model, which is algebraically equivalent to the model of Goldestein's. Thum (1997) presented a two-stage multivariate hierarchical linear model, which is the simplification of any three-stage HLM. Other scholars (e.g. Raudenbush et al, 1991) also formulated the multivariate multilevel model for the analysis of repeated measurement data.

These models all are appropriate for continuous dependent variables. However, Goldestein's multivariate multilevel linear model was also extended to analyze binary response data (Hox, 2002; Griffiths, 2004; Mcleod, 2001; Yang et al 2000). They set up two types of the Multivariate Multilevel Logistic Regression model (MMLR). Hox (2002: p.161-166) set one kind of MMLR model. The others showed the different model from his. Mcleod (2001) presented the MMLR model and gave the details how to extend Goldstein's multivariate multilevel model to binary dependent variables. Yang et al (2000) applied the univariate and multivariate multilevel logistic regression models to analyze the repeated binary outcomes for attitudes and voting over the electoral cycle. Griffiths et al (2004) compared univariate and multivariate logistic regression models for repeated measures for analysis of antenatal care in Uttar Pradesh. There are also some other multivariate logistic models which were not created within the multilevel framework (e.g. Glonek & McCullagh, 1995; Agresti, 1997). This dissertation will introduce Mcleod's MMLR model to the detection of DIF.

3.2 Multivariate Multilevel Linear Models

In Goldstein's multivariate multilevel linear model, a dummy variable is used to differentiate the different dependent variables and there is no random effect at level-1. Random effects for different dependent variables are put at level-2.

Suppose that we have *n* realizations of a multivariate random vector ***y*** of dimension *k*, i.e., *k* measurements or observations made on each of *n* individuals, and ***y*** is assumed to conform to a multivariate normal distribution (MVN). In light of the description of Mcleod (2001), the general level-1 model is written as:

$$y_{ij} = \sum_{t=1}^{p} z_{tij}(\beta_{0tj} + \sum_{q} \beta_{qtj} x_{qj})$$
$$where\ \ z_{tij} = \begin{cases} 1 \ \ for\ \ t = i \\ 0 \ \ otherwise \end{cases} \tag{2}$$

Thus y_{ij} is the i^{th} response for the j^{th} individual, and t indexes a set of k measurements. The zs are dummy variables used to distinguish between the k dependent variables. When $t{=}i$, the level-2 model is shown as follows:

$$\begin{aligned}
&\beta_{0ij} = \gamma_{0i} + e_{ij} \quad for\ i = 1, 2, \ldots, k. \\
&\beta_{qij} = \gamma_{qi} \\
&where\ e_j = \begin{bmatrix} e_{1j} \\ e_{2j} \\ \vdots \\ e_{kj} \end{bmatrix} \sim MVN\left(\vec{0}, \begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \cdots & & \ddots & \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_k^2 \end{bmatrix} \right)
\end{aligned} \qquad (3)$$

This model also can be extended to 3-level multivariate model if the individuals are nested in another unit.

As compared with the univariate multilevel model, this model has the following advantages: conclusions can be drawn about the correlations between the dependent variables (clearly the scores of individual items are correlated within persons as the same person provides responses to all items on a test); and the tests of specific effects for single dependent variables are more powerful in multivariate analysis (Snijders & Bosker, 1999).

3.3 Multivariate Multilevel Logistic Regression Models

The Multivariate Multilevel Logistic Regression Model (MMLR) is an extension of Goldstein's Multivariate Multilevel Model (Goldstein, 1995). In the MMLR model, the dependent variable y_{ij} is binary but is still the i^{th} response of individual j. Suppose that **y** is a vector and given the probability p_{ij} that the i^{th} response for the j^{th} individual, y_{ij} is assumed to conform to the Bernoulli distribution for any i and j, that is,

$y_{ij}|p_{ij} \sim Bernoulli(p_{ij})$. If the logit transformation is used, the generalized multivarite model is written by Mcleod (2001) as:

$$\log(\frac{p_{ij}}{1-p_{ij}}) = \sum_t [z_{tij}(\beta_{0tj} + \sum_q \beta_{qtj}x_{qj})] \qquad \textbf{(4)}$$

where the notation has the same meaning as before.

But how do we deal with random effects? Given p_{ij} the expected value of y_{ij} is p_{ij} and the variance is $p_{ij}(1-p_{ij})$. In terms of the explanations of Mcleod (2001) and Snijder and Bosker (1999), the following equations are given:

$$\begin{aligned} &y_{ij} = p_{ij} + e_{ij}\sqrt{p_{ij}(1-p_{ij})} \\ &where\ \ E(e_{ij}) = 0, Var(e_{ij}) = \sigma_i^2 \end{aligned} \qquad \textbf{(5)}$$

σ_i^2 is known as "extra-binomial variation", which is used to test the assumption of the Bernoulli distribution (Goldstein, 1995), and also named as the extra-binomial parameter (Yang et al, 2000), the extra-dispersion parameter (Guo & Zhao, 2000) or the scale parameter (Mcleod, 2001). If it is approximately equal to 1, y_{ij} conforms to the Bernoulli distribution. The variance of $\boldsymbol{e_j}$ has the following form (Yang et al, 2000; Griffiths et al, 2004):

$$Var(e_j) = \begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \cdots & & \ddots & \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_k^2 \end{bmatrix}. \qquad \textbf{(6)}$$

When the model is applied to solve some practical problem, different variance structures shown by Equation (6) can be assumed and used, such as diagonal, compound-symmetric, autoregressive (AR), autoregressive moving average (ARMA), unconstrained structure and so on.

In Hox's MMLR model, he simply used a logit transformation of p_{ij} in Equation (4) to take place of the dependent variable y_{ij} in Equation (2) (Hox, 2002: p.161-166).

3.4 MMLR DIF Detection Model

In the case for the detection of DIF, y_{ij} is person j's response to the i^{th} item. If G_j is a dummy variable for the group membership and G_j=1 when person j belongs to the focal group, otherwise G_j=0, a basic MMLR DIF detection model can be presented as the following equations:

$$\begin{aligned} \log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \sum_{t=1}^{k} z_{tij}(\beta_{0t} + \beta_{1t}G_j) \\ y_{ij} &= p_{ij} + e_{ij}\sqrt{p_{ij}(1-p_{ij})} \end{aligned} \qquad \textbf{(7)}$$

$$and\ E(e_j) = \vec{0}, Var(e_j) = \begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \cdots & & \ddots & \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_k^2 \end{bmatrix} \qquad \textbf{(8)}$$

where the zs are dummy variables used to distinguish between the individual's responses to different items; k is the number of items. If the examinees are nested in classes or schools, another level also could be added into the model. Here e_j is called the R-side random effect in SAS PROC GLIMMIX (SAS Institute, 2008), which is introduced with the G-side random effect in Chapter 5.

If Equation (2) is used to model random effects in MMLR, the model is not multivariate, but univariate. When the multivariate multilevel model was presented in the last section, there was no random effect for level-1. But for the multilevel logistic regression model, the level-1 model has a constant variance $\pi^{2/3}$ for the logit transformation (Snijder & Bosker, 1999) when the scale parameter is set to be 1. This implies there is variation between the different binary responses. So, if the multiple dependent variables have the same scale and measure the same thing, for instance, repeated measures, and they are thought to be nested in a level-2 unit, the variance matrix can be set as in Equation (2).

However, Equation (7) does not have a matching variable. Since "DIF is defined as item performance differences between examinees of comparable proficiency" (Camilli, 2007: p. 236), the MMLR model must include a matching variable to control the disparities of ability estimate between two groups. Suppose W_j is the ability estimate of the j^{th} person, then for MMLR, Equation (7) will be rewritten as follows:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \sum_{t=1}^{k} z_{tij}(\beta_{0t} + \beta_{1t}G_j + \beta_{2t}W_j)\,. \tag{9}$$

Now this model looks like a 2 Parameter Logistic (2PL) IRT model with DIF parameters. β_{t2} corresponds to the discrimination parameter, and $-(\beta_{0t}+ \beta_{1t}G_j)$ to the product of the discrimination and difficulty paramters. Or, in terms of the Rasch model, when the discrimination parameters of all items is set to the same value, the model also can be rewitten as:

$$\log(\frac{p_{ij}}{1-p_{ij}}) = \alpha W_j + \sum_{t=1}^{k} z_{tij}(\beta_{0t} + \beta_{1t} G_j) . \qquad \textbf{(10)}$$

Equation (10) can be regarded as the special case of Equation (9) when $\beta_{21}=\cdots=\beta_{2k}=\alpha$.

The variance matrix shown in Equation (8) makes it likely that MMLR can deal with the correlations between the responses of the same examinee to different items. For the DIF detection, however, it is only possible to select diagonal, compound-sysmetric or unconstrained structure. The type of autoregrssive structures are not reasonable since it is impossible to explain it in practice. In the simulation study, the simplest variance structure is assumed. First, $\sigma_{ij}=0$ ($i, j,$ = 1, 2, …, k and $i \neq j$) is set in the matrix of Equation (8). If $\sigma_{ij}\neq 0$, it means that the responses of an examinee to different items are correlated or locally dependent. It conflicts with the local item independence assumption. But the simulated data are generated based on the assumption. So, the constraint $\sigma_{ij}=0$ is set. Then, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \phi$ is also assumed to simplify the operation. Now, Var($\boldsymbol{e_j}$) $=\boldsymbol{\phi}\mathbf{I}$ where **I** stands for the identity matrix. SAS PROC GLIMMIX is able to be employed to estimate this type of models.

Chapter 4

HGLM DIF Detection Methods

This chapter describes the Hierarchical Generalized Linear Model DIF detection method, points out the confusing equations and potential problems, and modifies the model to solve these problems.

4.1 Kamta's HGLM DIF Model

The earlier papers about the multilevel models for binary data were published in 1985 (Guo & Zhou, 2000). This type of multilevel model is included in the Hierarchical Generalized Linear Model (HGLM) framework (Raudenbush & Bryk, 2002). Kamata (1998, 2001) outlined the extension of HGLM to IRT-style item analysis and the DIF analysis. His model assumes, given the item effects and the test-takers' abilities, y_{ij} takes on a value of 1 with probability p_{ij}. The level-1 model is

$$y_{ij} \big| p_{ij} \sim Bernoulli(p_{ij})$$

$$\eta_{ij} = \log\frac{p_{ij}}{1-p_{ij}} = \pi_{oj} + \sum_{q=1}^{k-1} \pi_{qj} z_{qij}, \tag{11}$$

where

z_{qij} is a dummy variable that takes on a value of unity if response i for person j is to item q, otherwise 0;

π_{qj} is thus the difference in log-odds of a correct response between item q and a "reference item" for examinee j;

k is the number of items.

While there are *k* items, only (*k*–1) dummy variables are included. The item not included is the reference item, whose difficulty is arbitrarily set to zero.

Then, an unconditional model is formulated for the abilities and all the item effects at level-2 model are fixed.

$$\begin{aligned} \pi_{0j} &= \beta_{00} + u_{0j}, \quad u_{0j} \sim N(0, \tau_{00}) \\ \pi_{qj} &= \beta_{q0} \quad for \quad q > 0 \end{aligned} \tag{12}$$

where u_{0j} is a random component of π_{0j} and it is distributed as a normal distribution with the mean of 0 and variance of τ_{00}. According to the studies of Kamata (1998; 2001), u_{0j} is considered to be the ability of person *j*, which is consistent with the one from BILOG based on the Rasch model, and $-(\beta_{q0}+\beta_{00})$ is correspondent to the difficulty of the Rasch model. Now the ability of persons is a random variable. In SAS PROC GLIMMIX, u_{0j} is the G-side random effect (SAS Institute Inc., 2008).

When HGLM model is used to find DIF, especially uniform DIF, the level-2 model changed into the one shown as follows:

$$\begin{aligned} \pi_{0j} &= \beta_{00} + u_{0j}, \quad u_{0j} \sim N(0, \tau_{00}) \\ \pi_{qj} &= \beta_{q0} + \beta_{q1} G_j \quad for \quad q > 0 \end{aligned} \tag{13}$$

where G_j denotes the group membership of person *j*. If the fixed coefficient of an item, β_{q1}, is significant, the item is thought to have uniform DIF.

However, this model does not examine whether the reference item has DIF or not if no group membership variable appears in the random intercept π_{0j} in Equation (13).

About this problem, Kamata's position causes some confusion. Sometimes he (e.g. Kamata, 1998; Kamata et al, 2005) put a group membership variable in the random intercept π_{0j} of Equation (13) to show if the reference item has DIF. It is feasible to do it theoretically and practically. But sometimes he (e.g. Kamata & Binici, 2003) deleted it from this random intercept π_{0j}. Besides, other researchers, such as Kim (2003) and Luppescu (2002), also gave the same model as Kamata and Binici did in their article of 2003. The likely reason is that they met a problem when interpreting the reference item with DIF. According to the definition of Raudenbush and Bryk (2002), π_{0j} is the ability of examinee j, so β_{00} should be the average ability of all examinees. Therefore, adding a group membership variable for π_{0j} can test if the examinees in the different groups have the different average ability. Of course, β_{00} can also be viewed as the difficulty of the reference item. But if its difficulty varies for the different groups, how can it be a reference?

4.2 The Modification of Kamata's HGLM

To reduce the difficulty of interpreting it, by using the notation of equations (11), (12), and (13), Kamata's unconditional model is reformulated as follows:

$$Level-1:$$

$$\begin{aligned} & y_{ij} | p_{ij} \sim Bernoulli(p_{ij}) \\ & \eta_{ij} = \log \frac{p_{ij}}{1-p_{ij}} = \pi_{oj} + \sum_{q=1}^{k} \pi_{qj} z_{qij} \end{aligned} \qquad \textbf{(14)}$$

$$Level-2:$$

$$\pi_{0j} = u_{0j}, \quad u_{0j} \sim N(0, \tau_{00})$$
$$\pi_{qj} = \beta_{q0} \quad for \quad q > 0$$

As compared with Equation (9), z_{qij} has the same meaning but π_{qj} or β_{q0} is the log-odds of person *j*'s response to item *q* now since no reference item is defined in the model. The unconditional model is still algebraically equivalent to Kamata's model, i.e., his generalized Rasch Model with random person ability (Kamata, 1998; 2001). The random intercept π_{0j} represents person ability and - β_{q0} are still correspondent to the estimate of item's difficulty.

Equation (14) reparameterizes Equation (11). In Equation (11), if dummy variables for all items are used, the matrix of independent variables of the equation will not be invertible so the coefficient of one dummy variable is zeroed out and its relevant item is defined as the reference item, which is called reference parameterization by Giesbrecht and Gumpertz (2004). But the matrix can also be invertible after deleting the level-1 intercept in Equation (11) and keeping all dummy variables for all items. Then it is changed into Equation (14). So, they are algebraically equivalent, but Equation (14) has *k* dummy variables at level-1 and no reference item.

When the model shown by Equation (14) is applied to look for DIF, the level-2 models of Equation (14) may be rewritten as:

$$\pi_{0j} = u_{0j}, \quad u_{0j} \sim N(0, \tau_{00})$$
$$\pi_{qj} = \beta_{q0} + \beta_{q1} G_j \quad for \quad q > 0 \qquad \textbf{(15)}$$

Then this DIF model is also mathematically equivalent to the HGLM DIF model which Kamata and others presented in their papers in 1998 and 2005. But it is more easily interpreted than Kamata's. β_{q1} will be used to test if an item shows DIF. In this study, the new DIF models will be used to identify DIF items. In this study, the new model will be called as the HGLM DIF model or detection procedure, and the original model of Kamata will be Kamata's HGLM.

In contrast with the standard logistic regression DIF procedure, the HGLM approach has a disadvantage before and after being modified. The random variable u_{0j} in HGLM is correspondent to the matching variable in standard logistic regression DIF model. The random variable should be the matching variable in HGLM because a matching variable must be constructed to create comparable subsets of examinees in DIF techniques (Camilli, 2007). These HGLM procedures assume that the ability of persons conform to the same normal distribution, that is, $N(0, \tau_{00})$ as mentioned in the above equations. So, theoretically, the ability of all examinees should have the same expected value 0 and variance τ_{00} although it can be regarded as the estimate of a person ability in practice. The assumption is not always true. Practically, after the group membership variable is added into the models, the random ability or the residual u_{0j} in HGLM Equation (12) and (14) will change, and it is not the ability estimate any more, so it is not able to be a matching variable to adjust the abilities of different groups or match the two groups. The subsequent simulation study will give the evidence to support the conclusion.

Therefore, Kim (2002) set up a HGLM DIF detection procedure with a matching variable, which can identify uniform and nonuniform DIF. In his procedure, the matching variable is the estimate of person ability, which is the residual u_{0j} of Equations (12) as suggested by Kamata (1998, 2001) and mentioned in Section 4.1. But the random effect was used as fixed in his method, so the HGLM procedure also needs a fixed matching variable W_j. Then if his procedure is just used to look for uniform DIF, Equation (15) can be reformulated as follows:

$$\begin{aligned} \pi_{0j} &= u_{0j}, \quad u_{0j} \sim N(0, \tau_{00}) \\ \pi_{qj} &= \beta_{q0} + \beta_{q1}G_j + \beta_{q2}W_j \quad for \quad q > 0 \end{aligned} \tag{16}$$

In the model, the item parameters π_{qj} are changed according to the person ability.

In addition, in light of Rasch model, Equation (15) can be shown as:

$$\begin{aligned} \pi_{0j} &= \beta_{01}W_j + u_{0j}, \quad u_{0j} \sim N(0, \tau_{00}) \\ \pi_{qj} &= \beta_{q0} + \beta_{q1}G_j \quad for \quad q > 0 \end{aligned} \tag{17}$$

π_{0j} is still an estimate of person ability, and then the random error can be described as the error of the measurement of person ability.

4.3 Differences between MMLR and HGLM

Generally, there are the following differences between the MMLR model and HGLM.

- MMLR is really a multivariate analytic method and deals with multiple dependent variables if $\sigma_{ij} \neq 0$ $(i \neq j)$ in the covariance matrix of Equation (6) while HGLM uses univariate method to do that as mentioned, and it is actually a special case of Hox's MMLR model. When every coefficient in Equation (11) is set as random, then it is Hox's unconditional MMLR model. So, one random coefficient makes it change from the multivariate to the univariate model.

- MMLR can test whether the extra-binomial or scale parameter is equal to 1 or not, viz., whether the dependent variables conform to the Bernoulli distribution. But the parameter is constrained to 1 in HGLM, so it is not possible to test whether the assumption is reasonable.
- As mentioned before, MMLR has no random effect at the first level and has it at the second level (Yang et al, 2000) while HGLM also has the random effects on both of level 1 and level 2.
- According to the definition of Kamata (1998, 2001), the residuals u_{0j} could be regarded as an estimate of person ability which is consistent with the estimate from the BILOG program. However, MMLR would not give the residual or the estimate of ability in that way.
- Finally, MMLR may deal with multidimensional tests more easily than HGLM when they are used to detect DIF in a test and appropriate matching variables are applied, for example, a science test including biology and physics items. If HGLM is used, another dummy variable needs to be added into the model and indicate different dimensions (Kamata, 1998) while MMLR does not need an additional dummy variable to do that since it is a real multivariate analysis.

Chapter 5

Estimation Methods

In this chapter, the estimation methods of MMLR and HGLM are introduced. The first section of the chapter presents the linearization and integral approximation methods, and explains why the linearization-based methods are selected in the study. The second section introduces some estimation methods in SAS GLIMMIX procedure. Finally, some details about these estimation methods are given.

5.1 Linearization and Integral Approximation Methods

It is easy to estimate the LR models by Maximum Likelihood. They will be estimated by SAS PROC LOGISTIC in the study. It is much more difficult and complicated to estimate MMLR and HGLM by Maximum Likelihood. At first, the marginal distribution of *Y* is approximated. The relevant approaches can mainly be classified into two broad categories, linearization and integral approximation methods (Schabenberger & Pierce, 2002). A linearization method approximates the nonlinear mixed model by a Taylor series to arrive at a pseudo-model and integral approximation methods use quadrature or Monte Carlo integration to calculate the marginal distribution of the data and maximize its likelihood (Schabenberger & Pierce, 2002). The linearization methods subsume Pseudo-Likelihood (PL), Penalized or Predictive Quasi-Likelihood (PQL) and Marginal Quasi-Likelihood (MQL). PL is almost the same as PQL and MQL except that PL explicitly estimates the extra-binomial or extra-dispersion parameter (Guo & Zhao, 2000). The integral approximation methods include Laplace, quadrature and Markov Chain Monte Carlo methods.

These linearization-based methods have a relatively simple form of the linearized model, which typically can be fit using only the mean and variance in the linearized form. The methods can fit the models for which the joint distribution is difficult or impossible to ascertain and the ones with correlated errors, a large number of random effects, crossed random effects, and multiple types of subjects. However, the approaches include the absence of a true objective function for the overall optimization process and potentially biased estimates of the covariance parameters, especially for binary data (SAS Institute, 2008), and these approaches are such crude approximations that the fit statistics based on the likelihood (e.g. deviance, Akaike's and Bayesian Information Criterion) are not recommended for use (Hox, 2002: p.110; Snijders & Bosker, 1999; p.220).

In contrast with the linearization-based methods, integral approximation methods provide an actual objective function for optimization, which enables researchers to perform likelihood ratio tests among nested models and to compute likelihood-based fit statistics (SAS Institute, 2008). The integral approximation methods are also more accurate than the linearization methods, e.g. Laplace approximation is more precise than PQL and MQL (Raudenbush, Yang & Yosef, 2000). But integral approximation methods are difficult for accommodating crossed random effects, multiple subject effects, and complex R-side covariance structures. Integral approximation methods are practically feasible for a small number of random effects (SAS Institute, 2008).

In light of these discussions, the linearization methods have to be selected to estimate MMLR in this study because it has complex R-side covariance structures. So, SAS PROC GLIMMIX is used to estimate MMLR since it implements one linearization-based methods — Pseudo-Likelihood. For the purpose of the comparison, it is better to apply the same software to HGLM so HGLM is also estimated by the procedure.

5.2 SAS GLIMMIX Procedure

SAS PROC GLIMMIX implements Pseudo-Likelihood. As noted, PL is almost the same as PQL and MQL except that PL explicitly estimates the extra-binomial or extra-dispersion parameter (Guo & Zhao, 2000). Some scholars (e.g. Van den Noortgate et al, 2003) even say SAS GLIMMIX macro (procedure) employs PQL and MQL. For the purpose of keeping consistent, it is also used to estimate HGLM. Although PQL and MQL were found to have downward bias (Breslow & Clayton, 1993; Rodriguez & Goldman, 1995), SAS GLIMMIX macro (procedure) is likely to be adequate for most of the projects undertaken in social science (Guo & Zhao, 2000) and even the first-order MQL is able to give acceptable estimates for less extreme datasets (Goldstein & Rasbash, 1996).

According to the *SAS/STAT® User's Guide* (SAS Institute, 2008), the GLIMMIX procedure can use four linearization-based methods to estimate these models. They are RSPL, MSPL, RMPL and MMPL. The abbreviation "PL" stands for pseudo-likelihood techniques. The first letter determines whether estimation is based on a residual likelihood ("R") or a maximum likelihood ("M"). The second letter identifies the expansion locus for the underlying approximation. The expansion locus of the first-order Taylor series expansion is either the vector of random effects solutions ("S") or the mean of the random effects ("M"). The expansions are also referred to as the "S"ubject-specific and “M”arginal expansions. RSPL is the default estimation method of PROC GLIMMIX. Of them, RSPL, MSPL are correspondent to PQL and RMPL and MMPL are MQL.

In the process of parameter estimation, several optimization techniques can be selected in the SAS procedure. For the Generalized Linear Mixed Model, the default is Quasi-Newton Optimization. To get the convergent outputs, other techniques also are used, such as Newton-Raphson Optimization with Line Search, Newton-Raphson Ridge Optimization, Quasi-Newton Optimization, etc. The details about these optimization techniques are shown in the *SAS/STAT® User's Guide*. Newton-Raphson Ridge Optimization is the default for pseudo-likelihood estimation with binary data in the procedure (SAS Institute Inc., 2008).When the simulated data are analyzed, the four estimation methods and these optimization techniques are used in turn until the outputs converge and RMPL and MMPL are used first since the first-order MQL is the most stable between the first- and second-order MQL and PQL (Snijders & Bosker, 1999).

5.3 Pseudo-Likelihood Estimation Based on Linearization

In terms of Wolfinger and O'Connell (1993) and the *SAS/STAT® User's Guide* (SAS Institute Inc., 2008), here are some details about the pseudo-likelihood estimation based on linearization. Suppose **Y** is the (n × 1) vector and represents the observed data; and γ is a (r × 1) vector of random effects. A Generalized Linear Mixed Model assume that

$$E[Y \mid \gamma] = g^{-1}(\eta) = \mu$$

where $\eta = X\beta + Z\gamma$; $g(\cdot)$ is a differentiable monotonic link function and $g^{-1}(\cdot)$ stands for its inverse. The matrix **X** is a ($n \times k$) matrix of rank k, and **Z** is a ($n \times r$) design matrix for the random effects. The G-side random effects are assumed to be normally distributed with mean **0** and variance matrix **G**. The variance of the observations, conditional on the random effects, is

$$Var[Y \mid \gamma] = A^{1/2}RA^{1/2},$$

Where the matrix **A** is a diagonal matrix and contains the variance functions of the model and **R** is the variance-covariance matrix between the R-side random effects, which is composed of Equation (6) or (7) in the study. In Section 3.4, the MMLR models are assumed to have the simplified covariance matrix, and only have R-side random effects. Then, **R** = ϕ**I** where **I** is the identity matrix and ϕ is the scale parameter. If a model has G-side random effects only, the procedure models R = ϕ**I**, When the HGLM model is fit in the study by SAS PROC GLIMMIX, ϕ**=1** is set.

Then, a first-order Taylor series of μ about $\widetilde{\beta}$ and $\widetilde{\gamma}$ yields

$$g^{-1}(\eta) \cong g^{-1}(\widetilde{\eta}) + \widetilde{\Delta}X(\beta - \widetilde{\beta}) + \widetilde{\Delta}Z(\gamma - \widetilde{\gamma})$$

where $\widetilde{\Delta} = \left(\dfrac{\partial g^{-1}(\eta)}{\partial \eta}\right)_{\widetilde{\beta},\widetilde{\gamma}}$ is a diagonal matrix of derivatives of the conditional mean evaluated at the expansion locus. If the terms are rearranged, the expression is:

$$\widetilde{\Delta}^{-1}[\mu - g^{-1}(\widetilde{\eta})] + X\widetilde{\beta} + Z\widetilde{\gamma}) \cong X\beta + Z\gamma$$

Its left-hand side is the expected value, conditional on, of

$$\widetilde{\Delta}^{-1}[Y - g^{-1}(\widetilde{\eta})] + X\widetilde{\beta} + Z\widetilde{\gamma}) \equiv P$$

and $Var(P \mid \gamma) = \widetilde{\Delta}^{-1}A^{1/2}RA^{1/2}\widetilde{\Delta}^{-1}$.

The model can thus be considered as **P** = $X\beta + Z\gamma + \varepsilon$, which is a linear mixed model with pseudo-response P, fixed effects β, random effects γ, and $Var(\varepsilon) = Var\ (P \mid \gamma\ )$.

Now in the linear mixed pseudo-model, the marginal variance can be defined as

$$Var(\vartheta) = ZGZ' + \widetilde{\Delta}^{-1}A^{1/2}RA^{1/2}\widetilde{\Delta}^{-1}$$

where ϑ is the parameter vector consisting of all unknowns in G and R. It is assumed that ε has a normal distribution and P is known. Then, the maximum log pseudo-likelihood function $\boldsymbol{l}\ (\vartheta, \boldsymbol{P})$ and restricted log pseudo-likelihood function $\boldsymbol{l}_R(\vartheta, \boldsymbol{P})$ are respectively gotten based on this linearized model. The former is used in MSPL and MMPL, and the latter is in RSPL and RMPL.

The fixed effects parameters $\boldsymbol{\beta}$ are profiled from these expressions. The parameters in ϑ are estimated by the optimization techniques. The objective function for minimization will be $-2\boldsymbol{l}\ (\vartheta, \boldsymbol{P})$ or $-2\boldsymbol{l}_R(\vartheta, \boldsymbol{P})$. At convergence, the profiled parameters $\boldsymbol{\beta}$ are estimated and the random effects ϑ are predicted as:

$$\widetilde{\beta} = (X'Var(\vartheta)^{-1}X)^{-1}X'Var(\vartheta)^{-1}P$$

$$\widetilde{\gamma} = \hat{G}Z'Var(\vartheta)^{-1}[P - X(X'Var(\vartheta)^{-1}X)^{-1}X'Var(\vartheta)^{-1}P]\ .$$

With the statistics, the pseudo-response and error weights of the linearized model are recalculated and the objective function is minimized again. In the approximated linear model, the predictors $\widetilde{\gamma}$ are the estimated Best Linear Unbiased Predictions. This process will continue until the relative change between parameter estimates at two successive iterations is sufficiently small. (SAS Institute Inc., 2008)

Chapter 6

Simulation Study

This chapter is about the simulation study. The study compares the performances of seven models, i.e. Reduced Logistic Regression, MMLR with Equations (7) (9) and (10) and HGLM with Equations (15) (16) and (17), detecting the uniform DIF. Power and Type I error rate of the seven models are calculated. Two types of matching variables are respectively employed, total raw score and ability estimate based on the 3PL IRT model.

6.1 Simulated Data

In this study, the efficiencies of the MMLR and HGLM models to identify DIF will be compared using a simulation study. The study was designed on the basis of the study by Narayanan and Swaminathan (1996). But some factors in their study will be excluded in this study, such as the proportion of item contamination, the sample size ratio of reference and focal groups, because their study was too complex and it had 384 conditions! This study just focuses on the influence of sample size, test length, DIF effect size and the ability distributions of person in the reference and focal groups.

First, Simulated data will be generated by SAS PROC IML, based on the 3-parameter logistic (3PL) IRT model. None of LR, MMLR and HGLM fit the data generated by the 3PL model, and none has in an advantage in the comparison. If the 2PL or the Rasch model were used, the simulated data would only fit some of these models and the other model would be at a disadvantage.

The following 3PL model is applied to generate the simulated data:

$$P(y = 1) = c + (1 - c)/\{1 + exp[-a(\theta - b - d)]\}.$$

In the above equation, θ is the person ability, a is the discrimination or slope which has uniform distribution on the interval [2/3, 1.5], b is the item difficulty which has the standard normal distribution, and c is lower asymptote or pseudo-guessing parameter which has uniform distribution on the interval [0, 0.2]. d represents the difference of b, i.e., item difficulty difference for the reference and focal groups, which result in DIF for the selected items. If an item has no DIF, then d is 0. If an item is randomly selected to have DIF, then the three levels of d, 0.25, 0.50 and 0.75, is set, and these differences are generated to favor the reference group over the focal group. But parameters a and c will not change and parameter b will be determined for the reference and focal groups to generate different DIF effect sizes since this study is only concerned with uniform DIF. Ability θ of the reference group are set to be normally distributed with mean 0 and variance 1, i.e., N(0,1) while θ of the focal group conform respectively to N(0, 1), N(0.5, 1) and N(0, 9). Many related studies (e.g. Finch & French, 2007; Jodoin & Gierl, 2001; Kristjansson et al, 2005; Narayanan & Swaninathan, 1996) showed the influence of the ability distributions with unequal means on the power and Type I error rate of the DIF detection procedures, but so far no other study discuss the influence of the difference of their variances except for Bolt and Gierl (2006).

Different sample sizes are selected to show sample size's effect on the DIF detection. A variety of sample sizes are used to explore the performance of the two methods. But the sample size ratio between the reference and focal group is equivalent for every test. The percentage, 50%, is arbitrarily set for the convenience. Three sample size conditions are simulated: (a) 400 total (200 in each group), (b) 700 total (350 in each group), (c) 1,000 total (500 in each group). Finch and French (2007) used three sample sizes: 500, 750, and 1000. To show more obvious effect, the differences between three sample sizes are increased a little from 250 to 300.

Finally, different test lengths are selected to show their effects on DIF detection. A variety of length sizes are chosen to examine the performance of the two procedures. Three test length conditions are simulated: (a) 20 items, (b) 40 items, (c) 60 items in each test, which were used by Whitmore and Schumacker (1999).

Then the three levels of *d*, i.e. the difference of parameter *b,* are selected, that is, 0.25, 0.50 and 0.75. The author tried 0.3, 0.6 and 0.9, which is from the design of Hidalgo and Lόpez-Pina (2004), i.e., 0.30, 0.60, and 1.00. But the trial simulation showed that 0.6 of b difference is large enough to make these methods to find most of DIF items, and 0.9 is too large to make these methods to find all DIF items under some conditions and not to show the detection differences. So, the average of 0.6 and 0.9, that is, 0.75 is selected and then 0.5 and 0.25 are selected in accordance with the study of Bolt and Gierl (2006). These differences are generated to favor the reference group over the focal group. An item's DIF effect size of IRT models should be quantified in terms of the area between the generating item response functions (Narayanan & Swaminathan, 1996; Swaminathan & Rogers, 1990). As the 3PL IRT model is used and parameters *a* and *c* keeps constant, the DIF effect size is $(1 - c)$ times the difference of parameter b in light of the formula of Raju (1988; 1990). Since $0 \leq c \leq 0.2$, the DIF effect size is between 0.2 and 0.25 when d=0.25, the effect size is between 0.40 and 0.50 when d=0.50, and the effect size is between 0.6 and 0.75 when d=0.75.

So, they will form 81 different conditions when the different types of sample size, test length and DIF effect size are crossed completely. For every condition, 500 simulation tests, i.e., 500 datasets, will be generated. In every dataset, the proportion of item contamination is kept the same. Twenty percent of all items are selected randomly and are set as DIF items. So, tests with 20, 40, and 60 items respectively have 4, 8 and 12 DIF items.

Finally, the generated probability p of each person's response to each item is compared to a uniform random number, which has a uniform distribution on the interval [0, 1]. If the probability is greater than the random number, the response is assigned the value of 1; otherwise, 0. At the same time, the different levels of sample size, test length and DIF effect size are simulated.

6.2 Some Practical Issues of the Simulation Study

6.2.1 Reduced LR DIF Model

The interaction term in the LR model shown by Model (1) may adversely influence the power when only uniform DIF is present because one degree of freedom is unnecessarily lost (Swaminathan & Rogers, 1990). Therefore, since this study only explores the performances of these procedures detecting uniform DIF, a reduced logistic regression (RLR) DIF model is applied in the study which is given by:

$$\log(\frac{p_j}{1-p_j}) = \beta_0 + \beta_1 W_j + \beta_2 G_j$$

where W_j is the matching variable and G_j is the group membership variable.

The simulated data were analyzed by RLR, MMLR with Equations (7) (9) and (10) (MMLR 7, MMLR 9 and MMLR 10) and HGLM with Equations (15) (16) and (17) (HGLM 15, HGLM 16 and HGLM 17). If MMLR performs better than or as well as RLR under this condition, it may has more efficient since it can deal with all items at one analysis.

6.2.2 Matching Variables

Since it is convenient to get total raw score for every person, total scores were used as the matching variables *W*. In addition, the R package ***ltm*** (Rizopoulos, 2006) was employed to estimate the person ability based on the 3PL IRT model and the IRT ability estimate also was used as the matching variable. In the R package, parameter estimates of the IRT models are obtained under marginal maximum likelihood using the Gauss-Hermite quadrature rule and then ability estimates can be obtained using Empirical Bayes (EB). These EB estimates are good measures of the person ability when the number of items tends to the infinity.

Table 1: Comparison of Means and Variances for the Matching Variables

Ability Distribution of focal group	Comparison of Means (t test)		Equality of Variances (Folded F test)	
	Total Score	IRT Ability Estimate	Total Score	IRT Ability Estimate
N(0, 1)	25.14%	24.15%	1.39%	3.64%
N(5, 1)	97.73%	98.05%	4.53%	5.05%
N(0, 9)	10.72%	6.12%	100%	100%

The *t* test and the Folded *F* test were applied to compare respectively the means and variances of the total scores and ability estimates of the reference and focal groups by SAS PROC TTEST. The *t* tests were based on the results of Folded *F* test, i.e. the *t* tests were regular ones with pooled variance estimate if the two groups were shown to have the equal variances and they were from Satterthwaite's method if the variances were unequal. Table 1 displays the rates of the significant tests in the all tests. At the 0.05 level of significance, the folded *F* test was able to find the unequal variances 100% in the study, and the error rates ranged from 1.39% to 5.05% and were not obvious while the *t* test had the correct rates of 98% but its highest error rate was 25%.

6.2.3 Evaluation Indexes

The accuracy of the these DIF detection methods were evaluated under a variety of conditions by Power and Type I error rate for detecting uniform DIF. Power is defined as the probability that an item that has DIF will be identified. Type I error rate is the probability that an item is identified to have DIF and in fact, really does not have it. In the DIF analysis, power was defined as the proportion of times that DIF is correctly identified while Type I error rate was defined as the proportion of times that a non-DIF item was falsely identified (Kristjansson et al, 2005).

First, in MMLR HGLM or RLR, the coefficient of G_j for each item was tested at significant level 0.05, and then it was judged that the item would be a DIF item, if the test was significant. It was known which judgment was true or false since the real DIF items were simulated. Finally, power was equal to the number of the correct judgments divided by the total number of all DIF items; and Type I error rate was the number of the wrong judgments divided by the total number of all non-DIF items. All of these judgments and calculations were only relevant to significant statistical tests.

6.3 Results of the Simulation Study

When total score was used as the matching variable, for three MMLR models, the extra-binomial or scale parameters were greater than 0.90 and smaller than 1.04. These could be regarded approximately as 1. So, the assumption of the Bernoulli distribution was tenable and the simulated data were not overdispersed or underdispersed (Goldstein, 1995; Yang et al, 2000). For HGLM and MMLR models, all simulated datasets had the convergent results provided by RMPL. For every condition, power and Type I error rate were computed, the results are listed in Table 4-9.

At first, a multivariate analysis of variance (MANOVA) was implemented using Wilks' Lambda to test the effect of test length, sample size, b difference and ability distributions for the reference and focal groups, the results of which are displayed in Table 2. According to these results, different test lengths, sample sizes, b differences and ability distributions of the two groups had significant effect on the power and Type I error rates of these seven models. After the computed power and error rates in Table 4-9 were checked, it was found, approximately:

- The longer the test, the larger power and Type I error rates.
- The more examinees take the test, the larger power and Type I error rates.
- The greater b-parameter difference between the two groups, the larger power and Type I error rates.
- The more different the ability distributions are for the two groups, especially, the more heterogeneous variances, the greater power and Type I error rates.

In general, for any condition, Type I error also inflates when power increases. It is impossible to raise power and reduce Type I error at the same time.

Table 2: Outputs of Multivariate Analysis of Variance

Effect	Wilks' Lambda	*F*	Den. D.F.	Num. D.F.	*p*-value
Test Length	3.433x10^{-4}	11.351	28	6	0.003
Sample Size	2.958x10^{-7}	393.793	28	6	0.000
b Difference	5.603x10^{-9}	2862.574	28	6	0.000
Distribution	2.589x10^{-11}	42114.723	28	6	0.000
Test Length x Sample Size	1.168x10^{-4}	2.288	56	13.8	0.046
Test Length x b Dif.	6.724x10^{-5}	2.675	56	13.8	0.024
Test Length x Distribution	1.133x10^{-6}	8.101	56	13.8	0.000
Sample Size x b Dif.	7.339x10^{-9}	30.249	56	13.8	0.000
Sample Size x Distribution	8.053x10^{-11}	97.039	56	13.8	0.000
b Dif. x distribution	1.190x10^{-12}	287.266	56	13.8	0.000
Test Length x Sample Size x b Dif.	1.652x10^{-6}	1.655	112	32.7	0.050
Test Length x Sample Size x distribution	1.663x10^{-6}	1.653	112	32.7	0.050
Test Length x b Dif. x Distribution	4.990x10^{-7}	2.017	112	32.7	0.012
Sample Size x b Dif. x Distribution	4.571x10^{-11}	8.403	112	32.7	0.000

Table 3: Multiple Comparisons of Power and Type I Error Rates

		HGLM 15	HGLM 16	HGLM 17	MMLR 7	MMLR 9	MMLR 10
Power	RLR	0.09* 0.03	0.000 0.000	-0.005 0.002	0.05 0.03	-0.004* 0.001	-0.008* 0.002
	HGLM 15				-0.04* 0.002		
	HGLM 16					-0.004* 0.001	
	HGLM 17						-0.003* 0.0004
Type I Error rate	RLR	-0.161* 0.03	0.000 0.000	-0.008* 0.001	-0.20* 0.03	-0.004* -0.001	-0.010* -0.001
	HGLM 15				-0.04* 0.002		
	HGLM 16					-0.004* 0.001	
	HGLM 17						-0.002* 0.0004

Note: In every cell, the upper number is the average difference and the lower is its standard error. * means $p < 0.0056$.

In the calculated power and Type I error rates listed in Table 4-9, under the same condition, HGLM16 and 17 had the similar results with RLR, and MMLR 7, 9 and 10 respectively had larger power or Type I error rate than HGLM 15, 16 and 17. Then, the repeated measure analysis of variance (Wilks' Lambda test) was used to compare these calculated power and error rates of the seven models detecting uniform DIF. The results showed that there were some significant differences of power and Type I error rate between these seven methods (for power, $\Lambda = 0.08$, F=135.45, degrees of freedom were 6 and 75, and $p < 0.0001$; for Type I error rate, $\Lambda = 0.11$, F=100.52, degrees of freedom were 6 and 75, and $p < 0.0001$.). So, the paired *t* test was used to make multiple comparisons between the pair of methods and a Bonferroni correction was employed to control the familywise error. Only the pairs that are of interest were compared. Totally, nine comparisons were made, so the level of significance was $0.05/9 \approx 0.0056$. Table 3 shows paired test's results of these comparisons.

By Table 3, MMLR 7, 9 and 10 respectively had significantly larger power or Type I error rate than HGLM 15, 16 and 17. The reason is that the estimates of the correspondent coefficients in these paired models were very similar, the difference was less than 1%, and the standard errors of these coefficients of MMLR were smaller than the counterparts of HGLM. As compared with RLR, generally, MMLR 7 had slightly smaller power and obviously larger error rate; MMLR 9 and 10 models had significantly larger power and smaller error rate; HGLM 15 had significantly smaller power and larger error rate; HGLM 16 had very similar results with RLR; and HGLM 17 had slightly larger power with RLR but significantly larger error rate than RLR. Actually, RLR and HGLM 16 had very similar estimates for the correspondent coefficients and their standard error, so their power and Type I error rate was also exactly the same in terms of Table 3.

MMLR 7 and HGLM 15 performed worse than RLR on power and Type I error rate while the other methods performed better at least on power or error rate. When the calculated power and error rates were checked carefully, the reason is obvious. When the ability distributions of the two groups had different means, the two methods performed very badly. Table 4-9 displays the power and Type I error rates of the seven models when the two groups have different ability distributions. By Table 5 and 8, almost for every condition combination, MMLR 7 and HGLM 15 had smaller power and larger Type I error rates than the others. Table 10 shows the mean power and Type I error rates of the seven models. By this table, on the average, the power rates MMLR 7 and HGLM 15 were less than 0.2 while the other methods' power rates were about 0.5; the error rates of the two models were greater than 0.6 while the other methods' error rates were about 0.09. MMLR 7 performed as badly as HGLM 15 under the condition. MMLR 7 has no matching variable while HGLM 15 has a random matching variable. These results mean that the random matching variable performs just like no matching variable in a DIF detection method. Therefore, MMLR 7 and HGLM 15 are not appropriate when the groups have different ability means.

Table 4 and 7 respectively show power and Type I error rates under every condition when the two groups have the same ability distribution, and Table 6 and 9 respectively show them when the two groups have the ability distributions with different variances. When power in Table 4 and 6 are compared, if the same methods are used under the same condition, power in Table 6 almost always is smaller than power in Table

4. When Type I error rates in Table 7 and 9 are compared, if the same methods are used under the same condition, error rates in Table 9 almost always is greater than power in Table 7. When power in Table 6 and Type I error rates in Table 9 are compared, for the same method, sometimes Type I error rate is even larger than power under the same conditions.

Table 10 also shows that when the two groups have the ability distributions with different variances, the average of power was only 0.29-0.41 and Type I error rate average was as high as 0.20-0.29 while under the other conditions the power was greater than 0.5 and the error rate was less than 0.1. It seems that the heterogeneous variances of the ability distributions significantly reduce power rates of these DIF detection approaches and inflate their Type I error rates in terms of the output of MANOVA in Table 2. The outputs are consistent with the study of Bolt and Gierl (2006). In their study, the disparities of these power and Type I error rates have appeared under the condition of unequal variances, but they did not find the obvious effect. In their study, the variances of two groups' ability distribution were set as 1 and 2, so the difference of the two variances was small and then the effect was too small to be noticed.

Table 4: Power by Methods for the Same Ability Distributions

Test Length	Sample Size	b Dif.	RLR	HGLM 15	HGLM 16	HGLM 17	MMLR 7	MMLR 9	MMLR 10
20	400	0.25	0.139	0.133	0.139	0.137	0.162	0.138	0.138
20	400	0.50	0.373	0.392	0.373	0.371	0.439	0.375	0.372
20	400	0.75	0.667	0.727	0.667	0.663	0.756	0.668	0.663
20	700	0.25	0.216	0.209	0.216	0.212	0.248	0.219	0.213
20	700	0.50	0.571	0.624	0.571	0.569	0.668	0.571	0.569
20	700	0.75	0.840	0.892	0.840	0.833	0.915	0.840	0.833
20	1000	0.25	0.282	0.281	0.282	0.278	0.325	0.284	0.279
20	1000	0.50	0.694	0.747	0.694	0.692	0.780	0.695	0.693
20	1000	0.75	0.912	0.948	0.912	0.901	0.956	0.913	0.901
40	400	0.25	0.135	0.125	0.135	0.134	0.155	0.135	0.133
40	400	0.50	0.398	0.41	0.398	0.393	0.459	0.398	0.391
40	400	0.75	0.659	0.724	0.659	0.655	0.755	0.659	0.655
40	700	0.25	0.204	0.202	0.204	0.200	0.239	0.205	0.200
40	700	0.50	0.574	0.627	0.574	0.571	0.676	0.575	0.570
40	700	0.75	0.837	0.893	0.837	0.826	0.911	0.837	0.825
40	1000	0.25	0.276	0.275	0.276	0.273	0.317	0.277	0.273
40	1000	0.50	0.701	0.765	0.701	0.695	0.804	0.701	0.694
40	1000	0.75	0.914	0.955	0.914	0.900	0.965	0.914	0.900
60	400	0.25	0.129	0.124	0.129	0.129	0.150	0.128	0.128
60	400	0.50	0.386	0.404	0.385	0.381	0.450	0.384	0.381
60	400	0.75	0.649	0.704	0.649	0.643	0.740	0.649	0.642
60	700	0.25	0.203	0.203	0.203	0.201	0.246	0.203	0.201
60	700	0.50	0.571	0.623	0.571	0.571	0.667	0.572	0.571
60	700	0.75	0.845	0.902	0.845	0.834	0.919	0.845	0.834
60	1000	0.25	0.265	0.27	0.265	0.263	0.317	0.266	0.263
60	1000	0.50	0.694	0.763	0.694	0.684	0.797	0.695	0.684
60	1000	0.75	0.915	0.951	0.915	0.904	0.961	0.915	0.904

Table 5: Power by Methods for Different Means of Ability Distributions

Test Length	Sample Size	b Dif.	RLR	HGLM 15	HGLM 16	HGLM 17	MMLR 7	MMLR 9	MMLR 10
20	400	0.25	0.144	0.129	0.144	0.154	0.155	0.146	0.154
20	400	0.50	0.386	0.035	0.386	0.392	0.053	0.389	0.393
20	400	0.75	0.681	0.130	0.680	0.702	0.157	0.684	0.702
20	700	0.25	0.184	0.207	0.184	0.194	0.246	0.190	0.197
20	700	0.50	0.600	0.034	0.600	0.622	0.045	0.606	0.624
20	700	0.75	0.868	0.201	0.868	0.905	0.235	0.869	0.907
20	1000	0.25	0.267	0.281	0.267	0.279	0.325	0.273	0.281
20	1000	0.50	0.722	0.038	0.722	0.754	0.055	0.724	0.757
20	1000	0.75	0.942	0.286	0.942	0.968	0.334	0.943	0.969
40	400	0.25	0.124	0.134	0.124	0.138	0.165	0.125	0.138
40	400	0.50	0.371	0.041	0.371	0.383	0.056	0.372	0.383
40	400	0.75	0.661	0.126	0.661	0.699	0.155	0.662	0.698
40	700	0.25	0.186	0.196	0.186	0.204	0.235	0.187	0.206
40	700	0.50	0.575	0.041	0.575	0.598	0.054	0.578	0.599
40	700	0.75	0.848	0.202	0.848	0.896	0.241	0.849	0.896
40	1000	0.25	0.243	0.273	0.243	0.253	0.323	0.245	0.255
40	1000	0.50	0.713	0.033	0.713	0.759	0.043	0.715	0.760
40	1000	0.75	0.926	0.278	0.926	0.963	0.328	0.926	0.963
60	400	0.25	0.131	0.127	0.131	0.139	0.160	0.131	0.139
60	400	0.50	0.366	0.034	0.366	0.380	0.049	0.366	0.379
60	400	0.75	0.660	0.133	0.660	0.698	0.166	0.661	0.696
60	700	0.25	0.194	0.200	0.194	0.204	0.241	0.197	0.204
60	700	0.50	0.564	0.039	0.564	0.595	0.056	0.566	0.596
60	700	0.75	0.843	0.204	0.843	0.894	0.244	0.845	0.894
60	1000	0.25	0.251	0.277	0.251	0.264	0.324	0.254	0.265
60	1000	0.50	0.695	0.037	0.695	0.737	0.054	0.697	0.738
60	1000	0.75	0.927	0.282	0.927	0.965	0.328	0.927	0.965

Table 6: Power by Methods for Different Variances of Ability Distributions

Test	Sample			HGLM	HGLM	HGLM	MMLR	MMLR	MMLR
Length	Size	b Dif.	RLR	15	16	17	7	9	10
20	400	0.25	0.092	0.163	0.092	0.097	0.229	0.098	0.107
20	400	0.50	0.193	0.241	0.193	0.195	0.308	0.202	0.204
20	400	0.75	0.357	0.360	0.357	0.346	0.425	0.371	0.359
20	700	0.25	0.121	0.268	0.121	0.127	0.336	0.134	0.136
20	700	0.50	0.276	0.374	0.276	0.270	0.435	0.291	0.279
20	700	0.75	0.520	0.481	0.520	0.508	0.535	0.528	0.517
20	1000	0.25	0.149	0.310	0.149	0.155	0.373	0.158	0.167
20	1000	0.50	0.374	0.432	0.374	0.352	0.472	0.388	0.364
20	1000	0.75	0.607	0.543	0.607	0.601	0.587	0.618	0.609
40	400	0.25	0.083	0.162	0.083	0.089	0.215	0.087	0.094
40	400	0.50	0.180	0.232	0.180	0.171	0.292	0.186	0.175
40	400	0.75	0.338	0.339	0.338	0.330	0.402	0.349	0.337
40	700	0.25	0.114	0.270	0.114	0.119	0.334	0.120	0.125
40	700	0.50	0.267	0.352	0.267	0.256	0.411	0.277	0.265
40	700	0.75	0.501	0.465	0.501	0.487	0.527	0.508	0.494
40	1000	0.25	0.151	0.339	0.151	0.159	0.405	0.159	0.166
40	1000	0.50	0.355	0.430	0.355	0.345	0.483	0.372	0.351
40	1000	0.75	0.602	0.550	0.602	0.587	0.599	0.612	0.593
60	400	0.25	0.093	0.172	0.093	0.098	0.232	0.098	0.101
60	400	0.50	0.172	0.237	0.172	0.171	0.302	0.182	0.176
60	400	0.75	0.328	0.343	0.328	0.319	0.414	0.340	0.325
60	700	0.25	0.121	0.263	0.121	0.120	0.326	0.127	0.125
60	700	0.50	0.262	0.360	0.262	0.255	0.427	0.274	0.263
60	700	0.75	0.500	0.483	0.500	0.483	0.540	0.511	0.489
60	1000	0.25	0.142	0.339	0.142	0.151	0.407	0.151	0.159
60	1000	0.50	0.350	0.440	0.350	0.336	0.493	0.360	0.345
60	1000	0.75	0.602	0.532	0.602	0.580	0.576	0.608	0.588

Table 7: Type I Error Rates by Methods for the Same Ability Distributions

Test Length	Sample Size	b Dif.	RLR	HGLM 15	HGLM 16	HGLM 17	MMLR 7	MMLR 9	MMLR 10
20	400	0.25	0.055	0.039	0.055	0.056	0.051	0.056	0.056
20	400	0.50	0.069	0.040	0.069	0.07	0.052	0.070	0.070
20	400	0.75	0.092	0.036	0.092	0.092	0.049	0.093	0.092
20	700	0.25	0.059	0.040	0.059	0.058	0.054	0.060	0.059
20	700	0.50	0.088	0.034	0.088	0.086	0.049	0.088	0.086
20	700	0.75	0.127	0.037	0.127	0.125	0.050	0.128	0.126
20	1000	0.25	0.064	0.037	0.064	0.066	0.050	0.065	0.066
20	1000	0.50	0.096	0.040	0.096	0.093	0.053	0.097	0.093
20	1000	0.75	0.163	0.037	0.163	0.158	0.050	0.164	0.159
40	400	0.25	0.055	0.037	0.055	0.054	0.049	0.055	0.054
40	400	0.50	0.071	0.036	0.071	0.071	0.050	0.071	0.071
40	400	0.75	0.092	0.037	0.092	0.092	0.051	0.092	0.091
40	700	0.25	0.061	0.037	0.061	0.061	0.051	0.061	0.061
40	700	0.50	0.088	0.037	0.088	0.086	0.050	0.088	0.086
40	700	0.75	0.128	0.037	0.128	0.124	0.051	0.129	0.124
40	1000	0.25	0.064	0.036	0.064	0.063	0.051	0.064	0.063
40	1000	0.50	0.100	0.038	0.100	0.095	0.053	0.101	0.095
40	1000	0.75	0.163	0.035	0.163	0.158	0.049	0.164	0.158
60	400	0.25	0.054	0.034	0.054	0.054	0.047	0.053	0.054
60	400	0.50	0.069	0.035	0.069	0.068	0.048	0.069	0.068
60	400	0.75	0.093	0.036	0.093	0.091	0.050	0.093	0.090
60	700	0.25	0.061	0.038	0.061	0.061	0.052	0.062	0.060
60	700	0.50	0.084	0.037	0.084	0.082	0.053	0.084	0.081
60	700	0.75	0.128	0.035	0.128	0.125	0.049	0.128	0.125
60	1000	0.25	0.063	0.035	0.063	0.062	0.048	0.064	0.062
60	1000	0.50	0.100	0.036	0.100	0.097	0.052	0.101	0.097
60	1000	0.75	0.162	0.032	0.162	0.155	0.044	0.163	0.155

Table 8: Type I Error Rates by Methods for Different Means of Ability Distributions

Test	Sample			HGLM	HGLM	HGL	MMLR	MMLR	MMLR
Length	Size	b Dif.	RLR	15	16	M 17	7	9	10
20	400	0.25	0.057	0.425	0.057	0.071	0.471	0.058	0.072
20	400	0.50	0.075	0.421	0.075	0.084	0.467	0.077	0.084
20	400	0.75	0.097	0.412	0.097	0.109	0.460	0.099	0.110
20	700	0.25	0.058	0.643	0.058	0.085	0.680	0.060	0.087
20	700	0.50	0.095	0.644	0.095	0.118	0.688	0.097	0.120
20	700	0.75	0.138	0.646	0.137	0.155	0.693	0.140	0.158
20	1000	0.25	0.068	0.776	0.067	0.102	0.808	0.070	0.104
20	1000	0.50	0.117	0.785	0.117	0.144	0.813	0.119	0.145
20	1000	0.75	0.188	0.790	0.188	0.212	0.821	0.192	0.214
40	400	0.25	0.058	0.432	0.058	0.072	0.481	0.058	0.072
40	400	0.50	0.074	0.420	0.074	0.086	0.470	0.075	0.086
40	400	0.75	0.097	0.427	0.097	0.109	0.477	0.097	0.109
40	700	0.25	0.061	0.642	0.061	0.085	0.686	0.062	0.085
40	700	0.50	0.091	0.639	0.091	0.114	0.686	0.093	0.114
40	700	0.75	0.139	0.645	0.139	0.161	0.689	0.141	0.161
40	1000	0.25	0.067	0.771	0.067	0.102	0.805	0.068	0.103
40	1000	0.50	0.113	0.775	0.113	0.145	0.807	0.115	0.145
40	1000	0.75	0.182	0.775	0.182	0.213	0.807	0.183	0.214
60	400	0.25	0.057	0.418	0.057	0.072	0.464	0.058	0.072
60	400	0.50	0.074	0.416	0.074	0.086	0.467	0.074	0.086
60	400	0.75	0.097	0.414	0.097	0.110	0.466	0.097	0.109
60	700	0.25	0.058	0.638	0.058	0.086	0.684	0.059	0.087
60	700	0.50	0.093	0.632	0.093	0.115	0.680	0.094	0.115
60	700	0.75	0.138	0.645	0.138	0.156	0.689	0.139	0.157
60	1000	0.25	0.066	0.778	0.066	0.101	0.810	0.067	0.102
60	1000	0.50	0.108	0.777	0.108	0.144	0.811	0.109	0.145
60	1000	0.75	0.173	0.770	0.173	0.206	0.806	0.175	0.207

Table 9: Type I Error Rates by Methods for Different Variances of Ability Distributions

Test	Sample	b		HGLM	HGLM	HGLM	MMLR	MMLR	MMLR
Length	Size	Dif.	RLR	15	16	17	7	9	10
20	400	0.25	0.116	0.143	0.116	0.121	0.196	0.124	0.127
20	400	0.50	0.134	0.142	0.134	0.138	0.194	0.143	0.144
20	400	0.75	0.161	0.143	0.161	0.168	0.197	0.171	0.176
20	700	0.25	0.158	0.226	0.158	0.165	0.291	0.168	0.173
20	700	0.50	0.191	0.223	0.191	0.190	0.288	0.202	0.199
20	700	0.75	0.243	0.229	0.243	0.245	0.294	0.255	0.255
20	1000	0.25	0.202	0.296	0.202	0.210	0.364	0.213	0.216
20	1000	0.50	0.250	0.302	0.250	0.258	0.369	0.264	0.267
20	1000	0.75	0.310	0.309	0.310	0.307	0.381	0.320	0.316
40	400	0.25	0.115	0.140	0.115	0.122	0.197	0.121	0.127
40	400	0.50	0.133	0.130	0.133	0.137	0.187	0.140	0.142
40	400	0.75	0.156	0.143	0.156	0.160	0.197	0.163	0.166
40	700	0.25	0.167	0.224	0.167	0.170	0.292	0.175	0.176
40	700	0.50	0.195	0.225	0.195	0.197	0.293	0.204	0.205
40	700	0.75	0.240	0.221	0.240	0.240	0.290	0.249	0.246
40	1000	0.25	0.208	0.299	0.208	0.213	0.369	0.219	0.218
40	1000	0.50	0.262	0.303	0.262	0.262	0.375	0.272	0.268
40	1000	0.75	0.309	0.296	0.309	0.302	0.364	0.321	0.308
60	400	0.25	0.116	0.135	0.116	0.120	0.193	0.121	0.125
60	400	0.50	0.134	0.131	0.134	0.138	0.186	0.141	0.143
60	400	0.75	0.160	0.132	0.160	0.160	0.188	0.166	0.166
60	700	0.25	0.161	0.222	0.161	0.166	0.290	0.169	0.172
60	700	0.50	0.199	0.225	0.199	0.200	0.294	0.207	0.205
60	700	0.75	0.243	0.222	0.243	0.241	0.291	0.252	0.246
60	1000	0.25	0.212	0.295	0.212	0.217	0.370	0.221	0.225
60	1000	0.50	0.256	0.297	0.256	0.254	0.369	0.266	0.260
60	1000	0.75	0.315	0.302	0.315	0.308	0.375	0.325	0.315

Table 10: Power and Type I Error Rates by Method and Ability Distributions

		Ability Distributions of Focal Group			
	Methods	N(0, 1)	N(.5, 1)	N(0, 9)	Total
Power	RLR	0.520	0.521	0.291	0.444
	HGLM15	0.551	0.148	0.351	0.350
	HGLM16	0.520	0.521	0.291	0.444
	HGLM17	0.515	0.546	0.285	0.449
	MMLR7	0.584	0.179	0.410	0.391
	MMLR9	0.521	0.523	0.300	0.448
	MMLR10	0.515	0.546	0.293	0.451
Type I Error rate	RLR	0.091	0.098	0.198	0.129
	HGLM15	0.037	0.613	0.220	0.290
	HGLM16	0.091	0.098	0.198	0.129
	HGLM17	0.089	0.120	0.200	0.136
	MMLR7	0.050	0.655	0.285	0.330
	MMLR9	0.091	0.099	0.207	0.132
	MMLR10	0.089	0.121	0.207	0.139

Table 11: Similarity Rates (%) of Results of the Seven Models

Model	HGLM 15	HGLM 16	HGLM 17	MMLR 7	MMLR 9	MMLR 10
RLR	71.08	100.00	95.69	68.91	99.63	95.61
HGLM15		71.08	71.23	95.94	70.95	71.15
HGLM16			95.69	68.91	99.63	95.61
HGLM17				69.06	95.66	99.73
MMLR7					68.83	69.01
MMLR9						95.63

Finally, the similarity of the DIF detection results for the seven models was evaluated. By Table 11, RLR and HGLM 16 almost gave the identical results for all items. Only 67 results were different among 1,620,000 times of detection. The identical judgments were more than 99% between RLR and MMLR 9. Between HGLM 17 and MMLR 10, the same results also exceeded 99%. It may mean that RLR and MMLR 10 almost give the same results if RLR and MMLR 10 are used to detect DIF and MMLR 10 has the R-side variance matrix assumptions of this study. The similarity rate between

HGLM 15 and MMLR 7 was also about 96%. It also supports the conclusion that the random matching variable performs like no matching variable in the DIF methods.

However, when total score was used as a matching variable in HGLM, the SAS GLIMMIX procedure gave the warnings that the covariance matrix is a zero matrix and the variance of the random effect was ZERO in the output no matter what estimation methods or optimization techniques were used in the SAS procedure. It is not reasonable. Actually, when RMPL was used, the estimates of the fixed effects and their standard errors of HGLM16 were almost the same as the correspondent ones gotten by RLR individually.

Since the variance estimate was not reasonable while total score was used as the matching variable in HGLM 16 and 17, the IRT ability estimate was tried as the matching variable. However, the variance estimates of the random effect in HGLM 16 and 17 were still equal to ZERO. For MMLR 9 and 10, the scale parameter was between 0.89 and 1.52. The range was larger than the one for the scale parameter that total score was used as the matching variable. Table 12 displays power and Type I error rates for the different matching variables. According to this table of the results from the paired *t*-tests, it was found that power and Type I error rates were improved for RLR, HGLM 16, 17, and MMLR 9 and 10 after the IRT ability estimate replaced total score as the matching variable in those models. If the comparisons, which were analogous to the ones when total score was used as the matching variable, were made the similar results and conclusions were obtained except that Type I error rates of MMLR 10 and HGLM 17 were not significantly different. The detailed power, Type I error and similarity rates are listed in Appendix I and II and the analysis results are in Appendix III.

Table 12: Power and Type I Error Rates Comparisons for the Different Matching Variables

	Model	Matching Variable		Difference	
		Total score	IRT ability estimate	Mean	Std. Error
Power	RLR	0.444	0.450	-0.006**	0.014
	HGLM16	0.444	0.450	-0.006**	0.014
	HGLM17	0.448	0.456	-0.007**	0.015
	MMLR9	0.448	0.453	-0.005**	0.014
	MMLR10	0.451	0.459	-0.007**	0.016
Type I error rate	RLR	0.129	0.117	0.012**	0.011
	HGLM16	0.129	0.117	0.012**	0.011
	HGLM17	0.136	0.125	0.011**	0.011
	MMLR9	0.132	0.120	0.013**	0.012
	MMLR10	0.139	0.125	0.014**	0.014

Note: * means $0.01 < p < 0.05$ and ** means $p < 0.01$ in paired *t*-tests. Five comparisons are made respectively for power and Type I error rate. So, if a Bonferroni correction is used, the adjusted significant level should be 0.05/5=0.01.

Chapter 7

Real Test Study

In this chapter, a Michigan Educational Assessment Program (MEAP) test is analyzed using the seven models, RLR, MMLR 7, 9 and 10, HGLM 15, 16 and 17. The uniform DIF items in the test are identified by these models. Two types of matching variables are respectively employed, total score and ability estimate based on the 3PL IRT model. The multilevel models are extended to have three levels.

7.1 Data

The data used in the study are from Office of Educational Assessment and Accountability, Michigan Department of Education. They are the third-graders' scores on Michigan Education Assessment Program (MEAP) reading test from fall of 2006. The study just analyzed 29 dichotomous items of the test. In 749 school districts, 118,245 students took the test. When SAS PROC GLIMMIX was used to estimate the models, a numerical value used in the estimation process was larger than the largest one allowed by SAS computing memory and the procedure was not able to run. So, 5% of all districts, i.e. 38 districts, were randomly selected so the SAS procedure can run the 2- and 3-level HGLM and MMLR models. In the selected sample, there are 6,351 students from Grade 3, including 3,125 girls and 3,226 boys, at these districts.

The students' total scores of the reading test were calculated and their reading abilities were estimated based on the 3PL IRT model by R package *ltm*. Table 13 shows the means and standard deviation of total scores and ability estimate of girls and boys.

Girls had higher scores and ability estimates than boys. First, the fold *F* test was applied to compare the variances of total scores and ability estimates of the girls and boys. The results were very significant (for total score, $F = 1.22, p < 0.0001$, for ability estimate, $F=1.18, p < 0.0001$). Then, by Satterthwaite's *t* test, the girls and boys had significantly different means of total scores and ability estimates (for total score, $t = 7.59$, degree of freedom = 6321, $p < 0.0001$; for the IRT ability estimate, $t = 7.65$, degree of freedom = 6333, $p < 0.0001$). Therefore, by the simulation study, HGLM 15 and MMLR 7 are not suitable for these data. If the two models were applied, they flagged 25 items as DIF items. The results would not be reasonable.

Table 13: Means of the Matching Variables by Gender

Variable	Gender	N	Mean	Std. Dev.
Total Score	F	3125	21.75	5.06
	M	3226	20.74	5.58
Ability Estimate	F	3125	-0.05	0.81
	M	3226	-0.22	0.88

7.2 Results of Two-Level MMLR and HGLM DIF Detection Methods

Since MMLR 7 and HGLM 15 were not appropriate, RLR, MMLR 9 and 10, and HGLM 16 and 17 were used to look for DIF items in the reading test respectively with the matching variables, total score and IRT ability estimate. At first, for MMLR 9 and 10, the R-side variance matrix was assumed as diagonal, which is the same as the matrix used in the simulation study, that is, $\sigma_{ij}=0$ and Var $(e_j) = \boldsymbol{\phi}\mathbf{I}$.

By Table 14, 15, and 16, whether total score or IRT ability estimate was used as the matching variable in RLR, MMLR 9 and HGLM 16, they gave the same results for DIF detection when the significant level 0.05 was applied. They flagged Item 8, 9, 11, 12, 16, 20, 24 and 26 as potential DIF items. MMLR 10 and HGLM 17 also gave the same detection results. They identified Item 9, 10, 11, 14, 16, 20, 24, and 28 as DIF item. The differences were Item 8 and 28. Even for the same item, however, its *p* values from different methods with different matching variables were differential.

When comparisons were made between these results, RLR, MMLR 9 and HGLM 16 gave similar coefficient estimates if the same matching variable was used in them. For RLR and HGLM 16, even the standard errors of these coefficients were very similar. But their standard errors from MMLR 9 were smaller than the ones from the former two methods. As noted in Section 6.3, the same phenomenon was also found by checking the estimate of the coefficients and their standard errors in the simulation study.

Table 14: Results of the Reduce Logistic Regression DIF Detection Models

	Matching Variable			
Item	Total Score		Ability Estimate	
1	-0.107	0.088	-0.127	0.088
2	0.012	0.057	0.016	0.056
3	0.010	0.054	0.030	0.053
4	-0.028	0.065	-0.038	0.065
5	0.008	0.058	-0.066	0.063
6	-0.069	0.064	-0.069	0.063
7	0.026	0.054	0.032	0.054
8	-0.181*	0.086	-0.232**	0.088
9	0.246**	0.084	0.219**	0.084
10	0.150	0.100	0.099	0.102
11	-0.214**	0.056	-0.196**	0.055
12	-0.001	0.078	-0.038	0.080
13	0.007	0.082	-0.018	0.082
14	0.348**	0.082	0.324**	0.083
15	-0.003	0.058	0.006	0.057
16	0.311**	0.104	0.261*	0.108
17	0.070	0.069	0.060	0.069
18	-0.063	0.082	-0.099	0.083
19	0.112	0.073	0.091	0.073
20	-0.110*	0.054	-0.091	0.053
21	-0.015	0.079	-0.036	0.079
23	0.061	0.083	0.046	0.083
24	-0.163**	0.061	-0.155*	0.060
25	0.070	0.061	0.075	0.060
26	-0.185*	0.080	-0.204*	0.080
27	-0.020	0.063	-0.003	0.062
28	-0.105	0.054	-0.087	0.053
29	-0.050	0.067	-0.049	0.066
30	-0.074	0.073	-0.079	0.072

Note: The table displays the coefficients that are relevant to DIF detection in every analysis and their standard errors. In each cell, the first number is the estimate of the coefficient and the second is its standard error. * means $0.01 < p < 0.05$ and ** means $p < 0.01$.

Table 15: Results of the DIF Detection (the Matching Variable: Total Score)

Item	HGLM 16		HGLM 17		MMLR 9		MMLR 10	
1	-0.107	0.088	-0.045	0.083	-0.107	0.085	-0.045	0.080
2	0.012	0.057	0.003	0.057	0.012	0.055	0.003	0.055
3	0.010	0.054	-0.084	0.058	0.010	0.052	-0.084	0.056
4	-0.028	0.065	0.003	0.063	-0.028	0.064	0.003	0.061
5	0.008	0.058	0.024	0.057	0.008	0.057	0.024	0.055
6	-0.069	0.064	-0.053	0.062	-0.069	0.062	-0.053	0.060
7	0.026	0.054	-0.027	0.057	0.026	0.053	-0.027	0.055
8	-0.181*	0.086	-0.063	0.078	-0.181*	0.084	-0.063	0.075
9	0.246**	0.084	0.272**	0.081	0.246**	0.081	0.272**	0.078
10	0.150	0.100	0.223*	0.092	0.150	0.097	0.223**	0.089
11	-0.214**	0.056	-0.240**	0.057	-0.214**	0.055	-0.240**	0.055
12	-0.001	0.078	0.083	0.071	-0.001	0.076	0.083	0.069
13	0.007	0.082	0.028	0.080	0.007	0.080	0.028	0.077
14	0.348**	0.082	0.364**	0.080	0.348**	0.080	0.364**	0.077
15	-0.003	0.058	-0.050	0.060	-0.003	0.056	-0.050	0.058
16	0.311*	0.104	0.370**	0.095	0.311**	0.101	0.370**	0.091
17	0.070	0.069	0.069	0.069	0.070	0.067	0.069	0.066
18	-0.063	0.082	0.014	0.076	-0.063	0.079	0.014	0.073
19	0.112	0.073	0.148*	0.069	0.112	0.071	0.148*	0.066
20	-0.110*	0.054	-0.194**	0.058	-0.110*	0.053	-0.194**	0.055
21	-0.015	0.079	0.028	0.075	-0.015	0.076	0.028	0.072
23	0.061	0.083	0.111	0.079	0.061	0.081	0.111	0.076
24	-0.163**	0.061	-0.167**	0.061	-0.163**	0.060	-0.167**	0.059
25	0.070	0.061	0.060	0.062	0.070	0.060	0.060	0.060
26	-0.185*	0.080	-0.099	0.075	-0.185*	0.078	-0.099	0.072
27	-0.020	0.063	-0.085	0.066	-0.020	0.061	-0.085	0.063
28	-0.105	0.054	-0.178**	0.057	-0.105*	0.052	-0.178**	0.055
29	-0.050	0.067	-0.054	0.067	-0.050	0.065	-0.054	0.064
30	-0.074	0.073	-0.051	0.071	-0.074	0.071	-0.051	0.069
Var.	0		0		0.947		0.930	

Note: The table displays the coefficients that are relevant to DIF detection in every analysis and their standard errors. In each cell, the first number is the estimate of the coefficient and the second is its standard error. * means 0.01< p <0.05 and ** means $p < 0.01$. The last row is the variance estimate of the random effect of the model.

Table 16: Results of the DIF Detection (the Matching Variable: Ability Estimate)

Item	HGLM 16		HGLM 17		MMLR 9		MMLR 10	
1	-0.127	0.088	-0.042	0.082	-0.127	0.087	-0.042	0.079
2	0.016	0.056	-0.006	0.057	0.016	0.056	-0.006	0.055
3	0.030	0.053	-0.092	0.058	0.030	0.053	-0.092	0.056
4	-0.038	0.065	-0.002	0.062	-0.038	0.065	-0.002	0.060
5	-0.066	0.063	0.014	0.057	-0.066	0.063	0.014	0.055
6	-0.069	0.063	-0.057	0.062	-0.069	0.062	-0.057	0.059
7	0.032	0.054	-0.039	0.058	0.032	0.054	-0.039	0.056
8	-0.232**	0.088	-0.061	0.077	-0.232**	0.087	-0.061	0.074
9	0.219**	0.084	0.266**	0.080	0.219**	0.084	0.266**	0.077
10	0.099	0.102	0.220	0.091	0.099	0.102	0.220*	0.088
11	-0.196**	0.055	-0.250**	0.057	-0.196**	0.055	-0.250**	0.055
12	-0.038	0.080	0.080	0.071	-0.038	0.079	0.080	0.068
13	-0.018	0.082	0.029	0.079	-0.018	0.082	0.029	0.076
14	0.324**	0.083	0.356**	0.079	0.324**	0.082	0.356**	0.076
15	0.006	0.057	-0.056	0.060	0.006	0.057	-0.056	0.058
16	0.261**	0.108	0.363	0.093	0.261*	0.107	0.363**	0.090
17	0.060	0.069	0.065	0.068	0.060	0.068	0.065	0.066
18	-0.099	0.083	0.014	0.075	-0.099	0.082	0.014	0.072
19	0.091	0.073	0.143*	0.068	0.091	0.072	0.143*	0.066
20	-0.091	0.053	-0.202**	0.057	-0.091	0.053	-0.202**	0.055
21	-0.036	0.079	0.028	0.074	-0.036	0.078	0.028	0.072
23	0.046	0.083	0.109	0.078	0.046	0.082	0.109	0.075
24	-0.155**	0.060	-0.171**	0.061	-0.155*	0.060	-0.171**	0.058
25	0.075	0.060	0.054	0.061	0.075	0.060	0.054	0.059
26	-0.204**	0.080	-0.097	0.074	-0.204*	0.080	-0.097	0.071
27	-0.003	0.061	-0.086	0.065	-0.003	0.061	-0.086	0.063
28	-0.087	0.053	-0.188**	0.057	-0.087	0.053	-0.188**	0.055
29	-0.049	0.066	-0.056	0.066	-0.049	0.066	-0.056	0.064
30	-0.079	0.072	-0.051	0.070	-0.079	0.072	-0.051	0.068
Var.					0.989		0.929	

Note: The table displays the coefficients that are relevant to DIF detection in every analysis and their standard errors. In each cell, the first number is the estimate of the coefficient and the second is its standard error. * means 0.01< p <0.05 and ** means p < 0.01. The last row is the variance estimate of the random effect of the model.

Then, the R-side variance matrix for MMLR 9 and 10 was assumed to be compound-symmetric (CS), which means that $\sigma_{ij} = \sigma \neq 0$ for any $i \neq j$ and $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \phi$. But when the total score was used as the matching variable, no convergent result was reached. Table 16 displays the partial results of MMLR with IRT ability estimate as the matching variable. The identified items were the same as the ones shown in Table 15. But when a different variance matrix was used for MMLR 9, the estimates of coefficients and their standard errors were different from the ones with the diagonal matrix, and both of them seemed to tend small. For Item 24, however, the *p*-value was just smaller than 0.05 when the simple diagonal variance matrix was set; but it was smaller than 0.01 when the variance matrix was compound symmetric. However, for MMLR 10, estimated σ was close to zero, these estimates were still similar to the ones in Table 16. The extremely small σ may means that the local independence assumption is tenable for the MMLR 10. Finally, the unconstrained matrix was tried, but no convergent results were reached.

Table 17: Results of the MMLR DIF Detection Models with CS Variance Matrix

Item	MMLR 9		MMLR 10	
1	-0.124	0.086	-0.042	0.079
2	0.016	0.055	-0.006	0.055
3	0.030	0.052	-0.092	0.056
4	-0.037	0.064	-0.002	0.060
5	-0.064	0.062	0.014	0.055
6	-0.069	0.062	-0.057	0.059
7	0.032	0.053	-0.039	0.056
8	-0.228**	0.086	-0.061	0.074
9	0.221**	0.083	0.266**	0.077
10	0.102	0.100	0.220*	0.088
11	-0.196**	0.054	-0.250**	0.055
12	-0.035	0.078	0.080	0.068
13	-0.016	0.081	0.029	0.076
14	0.326**	0.081	0.356**	0.076
15	0.006	0.056	-0.056	0.058
16	0.263*	0.105	0.363**	0.090
17	0.061	0.068	0.065	0.066
18	-0.096	0.081	0.014	0.072
19	0.093	0.071	0.143*	0.066
20	-0.092	0.053	-0.202**	0.055
21	-0.034	0.077	0.028	0.072
23	0.048	0.081	0.109	0.075
24	-0.154**	0.060	-0.171**	0.058
25	0.075	0.059	0.054	0.059
26	-0.202*	0.079	-0.097	0.071
27	-0.003	0.061	-0.086	0.063
28	-0.088	0.052	-0.188**	0.055
29	-0.048	0.065	-0.056	0.064
30	-0.078	0.071	-0.051	0.068
ϕ	0.975		0.929	
σ	-.0101		$1.27\text{x}10^{-15}$	

Note: The table displays the coefficients that are relevant to DIF detection in every analysis and their standard errors. In each cell, the first number is the estimate of the coefficient and the second is its standard error. * means 0.01< p <0.05 and ** means p < 0.01. The last row is the variance estimate of the random effect of the model.

7.3 Results of Three-Level MMLR and HGLM DIF Detection Methods

Since students are nested in different districts, these 2-level HGLM and MMLR models are extended to have 3 levels. If y_{ijl} is the response of student *j* in district *l* to item *i*, and for *t*=*i*, z_{tijl}=1; otherwise z_{tijl}=0, in terms of Mcleod (2001), a simple 3-level extension for MMLR 9 (3L MMLR 9) could be written as follows:

$$\log(\frac{p_{ijl}}{1-p_{ijl}}) = \sum_{t=1}^{k} z_{tijl}(\beta_{0tl} + \beta_{1t}G_{jl} + \beta_{2t}W_{jl})$$

$$y_{ijl} = p_{ijl} + e_{ijl}\sqrt{p_{ijl}(1-p_{ijl})}$$

$$\textit{and } E(e_{jl}) = \vec{0}, Var(e_{jl}) = \begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \cdots & & \ddots & \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_k^2 \end{bmatrix} \textit{ for any j and l.}$$

The third level model is: β_{0il}=γ_{00i}+u_{0i} where u_{0il}~N(0, τ_{00l}) for any *l*, and

$$T_l = Var\begin{pmatrix} u_{01l} \\ u_{02l} \\ \vdots \\ u_{0kl} \end{pmatrix} = \begin{bmatrix} \tau_{001} & & & \\ \tau_{021} & \tau_{002} & & \\ \cdots & & \ddots & \\ \tau_{0k1} & \tau_{0k2} & \cdots & \tau_{00k} \end{bmatrix} \text{ for any } l.$$

For the 3-level extension of MMLR 10 (3L MMLR 10), the level-1 model is shown as follows and the others are the same as the ones for MMLR 9.

$$\log(\frac{p_{ijl}}{1-p_{ijl}}) = \alpha W_{jl} + \sum_{t=1}^{k} z_{tijl}(\beta_{0tl} + \beta_{1t}G_{jl})$$

Now the 3-level MMLR model has the random effects at both of R- and G- sides. The constraint Var (e_{jl}) =$\boldsymbol{\phi}$**I** for the R-side variance matrix is still set, and the same type of matrix also is set for the G-side variance matrix, namely, $T_l = \boldsymbol{\tau}\mathbf{I}$. Total score and IRT ability estimate were still respectively used as the matching variable in this real test study. The results are shown in Table 18.

Table 18 shows that $\boldsymbol{\phi}$ was between 0.9 and 1, so the assumption of the Bernoulli distribution was tenable. By Table 18, under the constraint, the same types of the 3-level MMLR models gave the same results of the DIF detection whether total score or IRT ability estimate was used as the matching variables. In contrast with the correspondent 2-level MMLR model, 3L MMLR 9 additionally identified Item 20 and 28 as DIF items while 3L MMLR 10 gave the same DIF-detection results as 2L MMLR 10.

However, practically, the unconstrained variance matrixes might be more reasonable than the diagonal ones. Unfortunately, SAS PROC GLIMMIX did not give any convergent output for the 3-Level MMLR models with the unconstrained variance matrixes at R- or G-sides.

Based on Equations (14) and (16), the 3-level HGLM DIF model is rewritten as follows:

Level-1:

$$y_{ijl} \big| p_{ijl} \sim \mathit{Bernoulli}\,(p_{ijl})$$

$$\eta_{ijl} = \log\frac{p_{ijl}}{1-p_{ijl}} = \pi_{ojl} + \sum_{q=1}^{k} \pi_{qjl} z_{qijl}$$

Level-2:

$$\pi_{0jl} = \beta_{00l} + u_{0jl}, \quad u_{0jl} \sim N(0, \tau_{00l})$$
$$\pi_{qj} = \beta_{q0l} + \beta_{q1l} G_{jl} + \beta_{q2} W_{jl} \quad for \quad q > 0$$

Level-3:

$$\beta_{00l} = r_{00l}, \quad r_{00l} \sim N(0, \tau_{000})$$
$$\beta_{qsl} = \gamma_{qsl} \quad for \quad q > 0; s = 0, 1, 2$$

If Equation (17) is used, then the leve-2 and leve-3 model can be shown as:

Level-2:

$$\pi_{0jl} = \beta_{00l} + \beta_{01l} W_{jl} + u_{0jl}, \quad u_{0jl} \sim N(0, \tau_{00l})$$
$$\pi_{qjl} = \beta_{q0l} + \beta_{q1l} G_{jl} \quad for \quad q > 0$$

Level-3:

$$\beta_{00l} = r_{00l}, \quad r_{00l} \sim N(0, \tau_{000})$$
$$\beta_{01l} = \gamma_{010}$$
$$\beta_{qsl} = \gamma_{qsl} \quad for \quad q > 0; s = 0, 1$$

The fixed part of the 3-level HGLM is the same as the one in HGLM 16 and 17. Because the estimates of variances at level-2 and -3 both were still zero, the results did not change.

Table 18: Results of the 3-Level MMLR DIF Detection Methods

	Matching Variable							
	Total Score				IRT Ability Estimate			
Item	MMLR 9		MMLR 10		MMLR 9		MMLR 10	
1	-0.103	0.084	-0.042	0.079	-0.122	0.086	-0.038	0.078
2	0.011	0.055	0.001	0.055	0.015	0.056	-0.008	0.055
3	0.014	0.052	-0.078	0.056	0.034	0.052	-0.087	0.056
4	-0.024	0.064	0.005	0.061	-0.035	0.065	0.001	0.060
5	0.004	0.056	0.018	0.054	-0.069	0.062	0.007	0.054
6	-0.069	0.062	-0.054	0.060	-0.070	0.062	-0.058	0.059
7	0.022	0.053	-0.031	0.055	0.028	0.053	-0.043	0.055
8	-0.182*	0.084	-0.063	0.075	-0.233**	0.087	-0.060	0.074
9	0.243**	0.082	0.269**	0.078	0.215*	0.084	0.262**	0.077
10	0.147	0.097	0.220*	0.089	0.096	0.102	0.218*	0.088
11	-0.210**	0.055	-0.232**	0.055	-0.193**	0.055	-0.242**	0.055
12	0.001	0.076	0.084	0.068	-0.036	0.079	0.081	0.067
13	0.006	0.080	0.029	0.077	-0.018	0.082	0.029	0.076
14	0.345**	0.080	0.362**	0.077	0.322**	0.082	0.353**	0.076
15	-0.007	0.056	-0.055	0.058	0.002	0.056	-0.061	0.057
16	0.315**	0.100	0.374**	0.091	0.266*	0.107	0.369**	0.089
17	0.070	0.067	0.072	0.066	0.060	0.068	0.068	0.065
18	-0.063	0.079	0.010	0.072	-0.099	0.082	0.010	0.072
19	0.113	0.070	0.148*	0.066	0.092	0.072	0.142	0.065
20	-0.112*	0.052	-0.195**	0.055	-0.093	0.053	-0.204**	0.055
21	-0.009	0.076	0.034	0.072	-0.029	0.078	0.035	0.071
23	0.057	0.081	0.105	0.076	0.043	0.082	0.104	0.075
24	-0.159**	0.059	-0.165**	0.059	-0.151*	0.060	-0.168**	0.058
25	0.070	0.060	0.061	0.060	0.074	0.060	0.054	0.059
26	-0.181*	0.079	-0.097	0.073	-0.201*	0.080	-0.094	0.072
27	-0.023	0.061	-0.081	0.063	-0.006	0.061	-0.083	0.063
28	-0.109*	0.052	-0.176**	0.055	-0.092	0.053	-0.187**	0.055
29	-0.051	0.065	-0.058	0.064	-0.050	0.065	-0.060	0.063
30	-0.075	0.072	-0.052	0.070	-0.080	0.073	-0.051	0.069
ϕ	0.940		0.922		0.983		0.921	
τ	0.038		0.045		0.034		0.045	

Note: The table displays the coefficients that are relevant to the DIF detection and their standard errors. In each cell, the first number is the estimate of the coefficient and the second is its standard error. * means $0.01 < p < 0.05$ and ** means $p < 0.01$. The last row is the variance estimate of the random effect of the model.

Chapter 8

Discussions

This chapter summarizes the findings of the study, shows the advantages and disadvantages of the MMLR DIF detection methods, illustrates the reasons that some results appear, and explains the limitations of the study.

8.1 Using of the MMLR DIF Detection Method

From the simulation study, the MMLR DIF models was shown to have greater power rate for DIF detection than RLR, and from the real test study, these model showed similar results. Although their Type I error rates of the DIF was also greater than RLR's, the similarity rate of the results between RLR and these MMLR models is greater than 95%, especially the rate of MMLR 9 is 99.6% when the diagonal variance matrix is applied. If the unstructured or other reasonable variance matrixes are employed, it is expected that MMLR will give more accurate results for the DIF detection. Then, if LR is able to be used to detect DIF, MMLR also should be able to be applied to detect DIF especially when large power is needed.

In contrast with other DIF detection methods, the main advantage is that MMLR is able to model the related items of a test. As a natural multilevel model, MMLR can include the variations of examinees from the different groups, such as classes, schools and districts. The standard logistic regression DIF model can identify nonuniform DIF, and so can MMLR if an interaction term between the group membership and the matching variables is put in Equation (9). Then Equation (9) is rewritten as follows:

$$\log(\frac{p_{ij}}{1-p_{ij}}) = \sum_{t=1}^{k} z_{tij}(\beta_{t0} + \beta_{t1}G_j + \beta_{t2}W_j + \beta_{t3}G_jW_j)$$

If β_{t3} is significantly different from zero, then the item has nonuniform DIF. By a similar process, the 3-Level MMLR nonuniform DIF model is also developed.

Owing to the limitations of the estimation software, MMLR is not appropriate when a great number of examinees take a test or a test has too many items. Even when the sample size is small and the test does not have too many items, the computer still takes much more time and resources to deal with MMLR than HGLM and LR. For the 2-level MMLR model, different estimates of the R-side variance matrix only seem to influence the standard errors of the coefficients when the matrix is diagonal, i. e. σ_{ij}=0 (i, j = 1, 2, ..., k and $i \neq j$). But the convergence is another problem if complex variance matrixes are applied.

When a test has k items, $k(k+1)$ parameters need to be estimated in the matrix. In the real test study, the MEAP reading test has 29 items, and then 406 parameters need estimating if an unconstrained matrix is assumed, and over a hundred thousand students took the test. It is a task impossible for SAS PROC GLIMMIX to estimate the MMLR model with the unconstrained variance matrix and thousands of examinees.

As shown by the results of this study, MMLR inflates Type I error rate (see Table 7, 8 and 9, and Appendix II) if the sample size and DIF effect size are large. The inflated Type I error rate may result from the pseudo-guessing parameter. Jodoin and Gierl (2003) suggested reducing Type I error rate of LR using an effect size measure of LR. Possibly, the additional statistic also needs to be developed for MMLR in the future when it is used to analyze the 3PL-model-fitted data.

Due to the characteristics of MMLR, in light of the study, there are some tips or recommendations for using MMLR to detect DIF.

First, care is needed when selecting the appropriate variance matrices for MMLR. The diagonal R-side variance matrix for MMLR is simple and helpful to save time to estimate the model. It can be used only when it is known that the local independent assumption is tenable. Some methods should be employed to measure local item dependence, which were discussed by Yen and Fitzpatrich (2006: p. 141), before using the variance matrix. When it is not sure whether the assumption is reasonable, the unconstrained variance matrix should be used if a test is not too long, or the compound-symmetric variance matrix if the test does have many items.

Second, the IRT ability estimate should be applied as the matching variable in MMLR instead of total score, if the estimate is available. The reason is that the simulation study shows that MMLR had larger power and smaller Type I error rates when the IRT ability estimate was used as the matching variable than when total score was used in the simulation study.

Third, MMLR model with Equation (9) can be used to detect nonuniform and uniform DIF, so the interaction term should be included when it is applied in case the nonuniform DIF is omitted.

Finally, A third level should be included if it is known that the examinees are nested within some clusters, such as schools, states or others, and these data are available.

The hierarchical structure of MMLR is its basic advantage, and the use of the third level may improve the estimation of the model and DIF detection.

8.2 HGLM Is Unsuitable for DIF Detection

By the simulation study in Chapter 6 and the analysis in Section 4.2, HGLM is not able to identify DIF until it has a fixed matching variable. However, when total score or IRT ability estimate are added into the HGLM DIF models as the matching variable, the variance estimate of the level-2 random effect is zero in the simulation and the real test studies. The variance estimate is not reasonable. It means that neither of the matching variables should be included in the model, or the random effect should be excluded. If the two variables are inappropriate for the HGLMs, then it is difficult to find a matching variable. If the random effect is excluded, then the model will be regular logistic regression model.

Why is the variance estimate zero? It may be because total score or any ability estimate is highly correlated to means of the independent variable y or η across the level-2 units in Equation (11) or (14). Then for the whole HGLM DIF models, most of variation is explained by the matching variable. So, the residual is so small that it is close to zero. Finally, SAS PROC GLIMMIX sets it as zero.

If the hypothesis is correct, then any matching variable might be highly correlated to the means across the level-2 units, i.e. examinees. Even if it is not correct, neither the total score nor IRT ability estimate is good as the matching variable. The method of Kim (2003), using the ability estimate from the residual u_{0j} in Equation (12) as a matching variable, was also tried when analyzing the MEAP data, but the variance estimate was still zero. It is a mystery why I am not able to replicate the study of Kim (2003). Then it is very difficult to find an appropriate matching variable for the HGLMs but the model must have one if we want to use it to find DIF. It is a dilemma. So, HGLM may not be suitable to identify DIF.

8.3 Effects of the Heterogeneous Variances on the DIF Detection Methods

In the study, it is found that the different variances of the ability distributions of the reference and focal groups influence power and Type I error rates of LR, HGLM and MMLR DIF approaches. There are two explanations. If the variance of one group's ability distribution is much different from the one of the other group's, it may make one group have more persons with extreme ability than the other. When these persons are matched in the DIF methods by a matching variable, sometimes the regular values of one group are matched with the outliers of the other group, and then the poor results appear.

Of course, for the HGLM models, the random matching variable for all persons are assumed to conform to the same normal distribution. The assumption is not tenable no matter whether the models have the group membership variable or not when the ability distributions of the two groups have different variances. It must influence the results of the DIF detection procedures.

8.4 Comparisons of the Coefficients and Their Errors in MMLR and HGLM

The simulation study shows if the three types of MMLR models, MMLR 7, 9 and 10 respectively correspondent to HGLM 15, 16 and 17, the estimate of the correspondent coefficients in these paired models are very similar, and their standard errors in MMLR are smaller than the counterparts of correspondent HGLM. Why does that happen? It is related to the specification and estimation of the two kinds of models.

For MMLR 7, combining the two parts of Equation (7), and given independent variables, the following equation is given:

$$y_{ij} = \{1 + Exp[-\sum_{t=1}^{k} z_{tij}(\beta_{0t} + \beta_{1t}G_j)]\}^{-1}$$
$$+ e_{ij}\sqrt{\{1 + \exp[-\sum_{t=1}^{k} z_{tij}(\beta_{0t} + \beta_{1t}G_j)]\}^{-2} \exp[-\sum_{t=1}^{k} z_{tij}(\beta_{0t} + \beta_{1t}G_j)]}$$

Then the expected value of y_{ij},

$$E(y_{ij} \mid X) = \{1 + \exp[-\sum_{t=1}^{k} z_{tij}(\beta_{0t} + \beta_{1t}G_j)]\}^{-1} \qquad (18)$$

where X is used to denote all independent variables in the models.

For the modified HGLM DIF procedure, the outcome variable y_{ij} can be written as $y_{ij}=p_{ij}+e_{ij}$, where $E(e_{ij})=0$ and $Var(e_{ij})=p_{ij}(1+p_{ij})$ (Snijders & Bosker, 1999). Then combining equations (12) and (14), y_{ij} can be written as follows:

$$y_{ij} = \{1 + \exp[-\sum_{q=1}^{k} z_{qij}(\beta_{q0} + \beta_{q1}G_j) - u_{oj}]\}^{-1} + e_{ij}$$

Since $E(e_{ij})=0$, if the first-order Taylor series expansion is used, the approximation of the expected value of y_{ij},

$$E(y_{ij} \mid X) \approx \{1 + \exp[-\sum_{q=1}^{k} z_{qij}(\beta_{q0} + \beta_{q1}G_j) - E(u_{oj})]\}^{-1}$$

Then $E(y_{ij} \mid X) \approx \{1 + \exp[-\sum_{t=1}^{k} z_{qij}(\beta_{q0} + \beta_{q1}G_j)]\}^{-1}$ (19)

So, if we compare Equation (18) with (19), the expected values of the outcome variables in the MMLR 7 and HGLM 15 have the same expressions. So, it may imply that the estimates of fixed effects in the two models will be similar if the same data set is analyzed, the random effect of HGLM is omitted, and estimated the coefficients of the two models by some estimation approaches, in which the first Taylor series expansion is applied to get the approximation, such as PQL, MQL and PLs. The situation happens when RMPL or MMPL is used to estimate MMLR 7 and HGLM 15. The two methods are similar to MQL. The difference between PQL and MQL is that the Taylor series is expanded around the condition u_{0j}=0 in MQL while it is expanded around approximate posterior mode in PQL (Raudenbush & Bryk, 2002). Then, if MQL (RMPL or MMPL) is used to estimate HGLM then u_{0j}=0 is applied, and MMLR 7 has no random effect at G-side and then u_{0j}=0, the coefficients in HGLM 15 and MMLR 7 will be estimated based on the same equation. So, they have the similar estimates of the coefficients, which were shown in the results of the simulation study.

However, HGLM actually has the two random effects, u_{0j} at level-2 or G-side and a random effect at level-1 or R-side. Comparatively, MMLR 7 only has the R-side random effect. The scale parameter will influence the standard errors of the regression coefficients as it influences them in the Generalized Linear Model. But it has little effect when it is close to 1. Actually, the parameter approximates to 1 in the simulation study. So, the estimates of fixed effects in MMLR will have smaller standard errors. Since the estimates are similar and HGLM are estimated by RMPL in the simulation study, the hypotheses tests are significant more easily in MMLR 7 than in HGLM 15. Therefore, MMLR7 always has greater power and Type I error rate than HGLM 15 under the same condition.

By the same reasoning, MMLR 9 and HGLM 16, MMLR 10 and HGLM 17 also respectively have the same expressions for fixed effects and the estimates of these fixed effects in the former also have the smaller standard errors. So, if the appropriate matching variable is used in these models and the estimate of the variance of the level-2 random effect is not zero, the former may still have greater power and Type I error rate than the latter when RMPL is employed to estimate the latter.

In the study, the estimate of the G-side variance is zero in HGLM 16 and 17, and then the HGLM models are changed into the regular logistic regression model. It is shown as follows:

$$\log\frac{p_{ij}}{1-p_{ij}} = \sum_{q=1}^{k} z_{qij}(\beta_{q0} + \beta_{q1}G_j + \beta_{q2}W_j) \qquad \textbf{(20)}$$

$$\log\frac{p_{ij}}{1-p_{ij}} = \beta_{01}W_j + \sum_{q=1}^{k} z_{qij}(\beta_{q0} + \beta_{q1}G_j\ ) \qquad \textbf{(21)}$$

Equations (20) and (21) are the regular logistic regression models respectively corresponding to HGLM 16 and 17. In this situation, the coefficient estimates are still similar in the two models and the differences depend on their standard errors. Actually, under this condition, the standard error of a coefficient of MMLR is the counterpart of the regular logistic regression model multiplied by the corresponding σ_i, i.e., the square root of the diagonal entries in the variance matrix of the 2-level MMLR model. So, the standard errors of the coefficient estimates in MMLR are smaller than the ones in the regular logistic regression model when its scale parameters are less than 1 if both of them have the same fixed model. But most of the scale parameters of MMLR in the simulation study are smaller than 1. Therefore, for the most cases, the MMLR models have larger power and Type I error rates than the correspondent HGLM models.

However, the simulation and the real test studies show that HGLM 16 without random effect has the same estimates of the coefficients and their standard error as RLR. Why? Like standard logistic regression DIF model, RLR runs individual analysis for each item. If *k* dummy variables are used to identify these models for the different items, these individual RLR models are combined by these variables and merged into one model, and then it is HGLM16 without any random effect, and expressed as Equation (20). These models are the simple collection of the RLR models for all items. The estimates of coefficients and their standard errors in the combined model do not change (but their estimates are respectively from SAS PROC GLIMMIX and LOGISTIC so some small differences still exist between the correspondent estimates in the simulation study) although the individual RLR models are put together. So, by this way, LR DIF model also is able to analyze all items in a single run.

8.5 Limitations

In the simulation study, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \phi$ and σ_{ij}=0 (i, j = 1, 2, …, k and $i \neq j$) are assumed. But these assumptions may not be correct or reasonable for real tests. At the same time, when the covariances are constrained to 0, it may make the multivariate model become univariate and the multivariate model will lose the advantage in contrast with the univarite model.

This simulation study is not designed to explore the effects of the sample size ratio between the reference and focal groups and the proportion of item contamination. The sample size ratio between the reference and focal groups may influence DIF detection. If the proportion of items contaminated with DIF is set at a different percentage, the results may be different.

In some statistical tests of the simulation study, e.g., Wilks' Lambda and paired *t* tests, the calculated power and Type I error rates of the 7 models are the dependent variables. They may not be normally distributed. For Wilks' Lambda tests in MANOVA, the test of the heterogeneity of variances is not able to be implemented because the results of RLR and HGLM 16 are too similar. So, it is unknown if they have heterogeneous variance matrices. These factors have effects on the robustness of the statistical tests.

For the real data, an unconstrained R-side variance matrix may be more reasonable than others. As mentioned, the convergence is a big problem. Even if the convergent output exists, SAS PROC GLIMMIX will take long time, possibly several days, to get the output. Maybe other multilevel model software need to be tried, for example *MLwiN*.

Finally, this study is not involved with nonuniform DIF. In this Chapter, It is mentioned that MMLR 9 is able to be extended to identify nonuniform DIF. HGLM also has an extension for nonuniform DIF. As noted in Section 4.2, Kim (2003) extended Kamata's model to identify nonuniform DIF. Equation (16) is the reduced form of his model. His level-2 model is written as follows:

$$\pi_{0j} = u_{0j}, \quad u_{0j} \sim N(0, \tau_{00})$$
$$\pi_{qj} = \beta_{q0} + \beta_{q1}G_j + \beta_{q2}W_j + \beta_{q3}(G_j \times W_j)\ for \quad q > 0$$

If this model and the MMLR nonuniform DIF model in Section 8.1 are applied to identify uniform and nonuniform DIF in the 2005 MEAP reading test, they give the same results as the full logistic regression DIF model when all of them use the same matching variable, but the estimate variance of u_{0j} is still 0. If the different matching variables, total score and IRT ability estimate, are respectively used in these methods, they give different the results.

APPENDICES

Appendix I:

The Calculated Power Rates for LR, HGLM and MMLR

The following numbers are calculated when the IRT ability estimate is used as the matching variable:

Test Length	Sample Size	b Difference	Focal group ability distributions	RLR	HGLM16	HGLM17	MMLR9	MMLR10
20	400	0.25	N(0,1)	0.139	0.139	0.141	0.140	0.142
20	400	0.50	N(0,1)	0.374	0.374	0.372	0.373	0.373
20	400	0.75	N(0,1)	0.679	0.679	0.672	0.676	0.673
20	700	0.25	N(0,1)	0.215	0.215	0.209	0.215	0.209
20	700	0.50	N(0,1)	0.581	0.581	0.577	0.582	0.576
20	700	0.75	N(0,1)	0.854	0.854	0.842	0.855	0.842
20	1000	0.25	N(0,1)	0.279	0.279	0.278	0.284	0.281
20	1000	0.50	N(0,1)	0.697	0.697	0.695	0.699	0.696
20	1000	0.75	N(0,1)	0.920	0.920	0.909	0.920	0.909
20	400	0.25	N(0,9)	0.135	0.135	0.134	0.135	0.133
20	400	0.50	N(0,9)	0.401	0.401	0.399	0.399	0.396
20	400	0.75	N(0,9)	0.666	0.666	0.663	0.665	0.660
20	700	0.25	N(0,9)	0.203	0.203	0.199	0.203	0.198
20	700	0.50	N(0,9)	0.576	0.576	0.579	0.576	0.576
20	700	0.75	N(0,9)	0.847	0.847	0.833	0.847	0.833
20	1000	0.25	N(0,9)	0.278	0.278	0.278	0.279	0.277
20	1000	0.50	N(0,9)	0.706	0.706	0.700	0.706	0.698
20	1000	0.75	N(0,9)	0.921	0.921	0.903	0.921	0.902
20	400	0.25	N(5,1)	0.128	0.128	0.127	0.127	0.125
20	400	0.50	N(5,1)	0.389	0.389	0.386	0.387	0.382
20	400	0.75	N(5,1)	0.657	0.657	0.648	0.655	0.645
20	700	0.25	N(5,1)	0.202	0.202	0.201	0.202	0.199
20	700	0.50	N(5,1)	0.574	0.574	0.573	0.574	0.572
20	700	0.75	N(5,1)	0.851	0.851	0.841	0.851	0.840
20	1000	0.25	N(5,1)	0.265	0.265	0.265	0.265	0.265
20	1000	0.50	N(5,1)	0.697	0.697	0.686	0.698	0.686
20	1000	0.75	N(5,1)	0.920	0.920	0.907	0.920	0.906
40	400	0.25	N(0,1)	0.087	0.087	0.090	0.092	0.097
40	400	0.50	N(0,1)	0.193	0.193	0.193	0.200	0.206
40	400	0.75	N(0,1)	0.388	0.388	0.365	0.399	0.382
40	700	0.25	N(0,1)	0.112	0.112	0.119	0.118	0.130
40	700	0.50	N(0,1)	0.299	0.299	0.288	0.312	0.305
40	700	0.75	N(0,1)	0.558	0.558	0.546	0.568	0.562
40	1000	0.25	N(0,1)	0.136	0.136	0.133	0.145	0.147

40	1000	0.50	N(0,1)	0.399	0.399	0.370	0.416	0.395
40	1000	0.75	N(0,1)	0.653	0.653	0.646	0.664	0.656
40	400	0.25	N(0,9)	0.073	0.073	0.079	0.077	0.081
40	400	0.50	N(0,9)	0.183	0.183	0.173	0.186	0.174
40	400	0.75	N(0,9)	0.362	0.362	0.352	0.367	0.352
40	700	0.25	N(0,9)	0.101	0.101	0.105	0.106	0.109
40	700	0.50	N(0,9)	0.282	0.282	0.271	0.291	0.274
40	700	0.75	N(0,9)	0.538	0.538	0.538	0.546	0.540
40	1000	0.25	N(0,9)	0.131	0.131	0.137	0.139	0.140
40	1000	0.50	N(0,9)	0.376	0.376	0.369	0.385	0.373
40	1000	0.75	N(0,9)	0.646	0.646	0.637	0.655	0.640
40	400	0.25	N(5,1)	0.083	0.083	0.087	0.084	0.088
40	400	0.50	N(5,1)	0.178	0.178	0.172	0.181	0.172
40	400	0.75	N(5,1)	0.352	0.351	0.345	0.353	0.346
40	700	0.25	N(5,1)	0.104	0.104	0.105	0.107	0.106
40	700	0.50	N(5,1)	0.276	0.276	0.271	0.282	0.273
40	700	0.75	N(5,1)	0.537	0.537	0.533	0.542	0.534
40	1000	0.25	N(5,1)	0.122	0.122	0.129	0.127	0.142
40	1000	0.50	N(5,1)	0.367	0.367	0.330	0.371	0.368
40	1000	0.75	N(5,1)	0.651	0.651	0.623	0.657	0.647
60	400	0.25	N(0,1)	0.148	0.148	0.145	0.152	0.148
60	400	0.50	N(0,1)	0.384	0.384	0.390	0.385	0.391
60	400	0.75	N(0,1)	0.676	0.676	0.715	0.680	0.718
60	700	0.25	N(0,1)	0.198	0.198	0.192	0.202	0.193
60	700	0.50	N(0,1)	0.588	0.588	0.633	0.596	0.640
60	700	0.75	N(0,1)	0.854	0.854	0.915	0.854	0.915
60	1000	0.25	N(0,1)	0.269	0.269	0.287	0.275	0.290
60	1000	0.50	N(0,1)	0.708	0.708	0.768	0.714	0.770
60	1000	0.75	N(0,1)	0.931	0.931	0.971	0.932	0.970
60	400	0.25	N(0,9)	0.125	0.125	0.135	0.125	0.135
60	400	0.50	N(0,9)	0.376	0.376	0.388	0.376	0.385
60	400	0.75	N(0,9)	0.664	0.664	0.718	0.664	0.715
60	700	0.25	N(0,9)	0.190	0.190	0.202	0.190	0.202
60	700	0.50	N(0,9)	0.580	0.580	0.609	0.582	0.608
60	700	0.75	N(0,9)	0.850	0.850	0.906	0.850	0.906
60	1000	0.25	N(0,9)	0.253	0.253	0.260	0.255	0.260
60	1000	0.50	N(0,9)	0.712	0.712	0.773	0.714	0.774
60	1000	0.75	N(0,9)	0.925	0.925	0.969	0.925	0.969
60	400	0.25	N(5,1)	0.133	0.133	0.140	0.133	0.139
60	400	0.50	N(5,1)	0.376	0.376	0.386	0.375	0.384
60	400	0.75	N(5,1)	0.672	0.672	0.716	0.671	0.713
60	700	0.25	N(5,1)	0.199	0.199	0.209	0.200	0.209
60	700	0.50	N(5,1)	0.571	0.571	0.608	0.572	0.608
60	700	0.75	N(5,1)	0.849	0.849	0.904	0.850	0.904
60	1000	0.25	N(5,1)	0.258	0.258	0.273	0.260	0.273
60	1000	0.50	N(5,1)	0.703	0.703	0.753	0.704	0.752
60	1000	0.75	N(5,1)	0.928	0.928	0.969	0.928	0.969

Appendix II:

The Calculated Type I error Rates for LR, HGLM and MMLR

The following numbers are calculated when the IRT ability estimate is used as the matching variable:

Test Length	Sample Size	b Difference	Focal group ability distributions	RLR	HGLM16	HGLM17	MMLR9	MMLR10
20	400	0.25	N(0,1)	0.055	0.055	0.053	0.055	0.053
20	400	0.50	N(0,1)	0.068	0.068	0.067	0.067	0.067
20	400	0.75	N(0,1)	0.088	0.088	0.084	0.088	0.084
20	700	0.25	N(0,1)	0.061	0.061	0.058	0.062	0.058
20	700	0.50	N(0,1)	0.085	0.085	0.083	0.086	0.083
20	700	0.75	N(0,1)	0.125	0.125	0.120	0.126	0.120
20	1000	0.25	N(0,1)	0.062	0.062	0.064	0.063	0.064
20	1000	0.50	N(0,1)	0.096	0.096	0.092	0.098	0.092
20	1000	0.75	N(0,1)	0.156	0.156	0.148	0.157	0.149
20	400	0.25	N(0,9)	0.054	0.054	0.054	0.053	0.053
20	400	0.50	N(0,9)	0.070	0.070	0.069	0.069	0.069
20	400	0.75	N(0,9)	0.088	0.088	0.086	0.087	0.085
20	700	0.25	N(0,9)	0.060	0.060	0.060	0.061	0.059
20	700	0.50	N(0,9)	0.089	0.089	0.085	0.089	0.084
20	700	0.75	N(0,9)	0.123	0.123	0.119	0.123	0.118
20	1000	0.25	N(0,9)	0.064	0.064	0.063	0.064	0.063
20	1000	0.50	N(0,9)	0.098	0.098	0.093	0.098	0.092
20	1000	0.75	N(0,9)	0.161	0.161	0.151	0.161	0.151
20	400	0.25	N(5,1)	0.054	0.054	0.055	0.054	0.054
20	400	0.50	N(5,1)	0.067	0.067	0.067	0.066	0.065
20	400	0.75	N(5,1)	0.090	0.090	0.088	0.089	0.086
20	700	0.25	N(5,1)	0.061	0.061	0.061	0.061	0.060
20	700	0.50	N(5,1)	0.082	0.082	0.079	0.082	0.078
20	700	0.75	N(5,1)	0.121	0.121	0.118	0.121	0.117
20	1000	0.25	N(5,1)	0.063	0.063	0.060	0.063	0.060
20	1000	0.50	N(5,1)	0.099	0.099	0.093	0.099	0.093
20	1000	0.75	N(5,1)	0.155	0.155	0.148	0.155	0.147
40	400	0.25	N(0,1)	0.103	0.103	0.109	0.108	0.111
40	400	0.50	N(0,1)	0.122	0.122	0.126	0.128	0.127
40	400	0.75	N(0,1)	0.145	0.145	0.151	0.151	0.151
40	700	0.25	N(0,1)	0.135	0.135	0.144	0.143	0.146
40	700	0.50	N(0,1)	0.167	0.167	0.167	0.177	0.168
40	700	0.75	N(0,1)	0.215	0.215	0.217	0.224	0.217
40	1000	0.25	N(0,1)	0.173	0.173	0.178	0.183	0.178

40	1000	0.50	N(0,1)	0.222	0.222	0.225	0.232	0.225
40	1000	0.75	N(0,1)	0.281	0.281	0.274	0.293	0.272
40	400	0.25	N(0,9)	0.102	0.102	0.108	0.106	0.110
40	400	0.50	N(0,9)	0.122	0.122	0.127	0.126	0.127
40	400	0.75	N(0,9)	0.148	0.148	0.151	0.152	0.151
40	700	0.25	N(0,9)	0.142	0.142	0.145	0.148	0.148
40	700	0.50	N(0,9)	0.173	0.173	0.173	0.179	0.175
40	700	0.75	N(0,9)	0.219	0.219	0.213	0.225	0.215
40	1000	0.25	N(0,9)	0.175	0.175	0.175	0.182	0.178
40	1000	0.50	N(0,9)	0.229	0.229	0.222	0.237	0.226
40	1000	0.75	N(0,9)	0.282	0.282	0.262	0.289	0.265
40	400	0.25	N(5,1)	0.105	0.105	0.111	0.107	0.111
40	400	0.50	N(5,1)	0.124	0.124	0.126	0.127	0.126
40	400	0.75	N(5,1)	0.149	0.149	0.149	0.151	0.148
40	700	0.25	N(5,1)	0.138	0.138	0.144	0.142	0.146
40	700	0.50	N(5,1)	0.179	0.179	0.176	0.184	0.177
40	700	0.75	N(5,1)	0.219	0.219	0.213	0.224	0.214
40	1000	0.25	N(5,1)	0.179	0.179	0.196	0.185	0.182
40	1000	0.50	N(5,1)	0.224	0.224	0.240	0.230	0.219
40	1000	0.75	N(5,1)	0.282	0.282	0.292	0.288	0.269
60	400	0.25	N(0,1)	0.056	0.056	0.070	0.057	0.071
60	400	0.50	N(0,1)	0.072	0.072	0.083	0.073	0.084
60	400	0.75	N(0,1)	0.088	0.088	0.099	0.088	0.100
60	700	0.25	N(0,1)	0.052	0.052	0.084	0.054	0.086
60	700	0.50	N(0,1)	0.085	0.085	0.112	0.087	0.115
60	700	0.75	N(0,1)	0.114	0.114	0.141	0.116	0.143
60	1000	0.25	N(0,1)	0.061	0.061	0.100	0.064	0.102
60	1000	0.50	N(0,1)	0.100	0.100	0.136	0.102	0.137
60	1000	0.75	N(0,1)	0.151	0.151	0.192	0.155	0.193
60	400	0.25	N(0,9)	0.055	0.055	0.072	0.055	0.071
60	400	0.50	N(0,9)	0.070	0.070	0.082	0.070	0.082
60	400	0.75	N(0,9)	0.084	0.084	0.098	0.085	0.097
60	700	0.25	N(0,9)	0.058	0.058	0.085	0.058	0.085
60	700	0.50	N(0,9)	0.083	0.083	0.110	0.084	0.109
60	700	0.75	N(0,9)	0.118	0.118	0.144	0.119	0.144
60	1000	0.25	N(0,9)	0.064	0.064	0.101	0.064	0.101
60	1000	0.50	N(0,9)	0.100	0.100	0.137	0.102	0.137
60	1000	0.75	N(0,9)	0.153	0.153	0.190	0.153	0.190
60	400	0.25	N(5,1)	0.057	0.057	0.072	0.057	0.072
60	400	0.50	N(5,1)	0.070	0.070	0.084	0.070	0.083
60	400	0.75	N(5,1)	0.086	0.086	0.099	0.086	0.098
60	700	0.25	N(5,1)	0.055	0.055	0.085	0.055	0.085
60	700	0.50	N(5,1)	0.085	0.085	0.110	0.086	0.109
60	700	0.75	N(5,1)	0.120	0.120	0.142	0.120	0.142
60	1000	0.25	N(5,1)	0.063	0.063	0.099	0.064	0.099
60	1000	0.50	N(5,1)	0.096	0.096	0.137	0.096	0.137
60	1000	0.75	N(5,1)	0.144	0.144	0.183	0.145	0.183

Appendix III:

Results of Analyses with the IRT Ability Estimate

The results displayed here are obtained when the IRT ability estimate is used as the matching variable.

1. Output of multivariate analysis of variance for power and Type I error rates of LR, MMLR7, MMLR9, MMLR10, HGLM15, HGLM16 and HGLM17.

Effect	Wilks' Lambda	F	Den. D.F.	Num. D.F.	p-value
Test Length	1.565×10^{-5}	53.947	28	6	0.000
Sample Size	8.471×10^{-8}	736.038	28	6	0.000
b Difference	4.808×10^{-9}	3090.202	28	6	0.000
Distribution	1.087×10^{-11}	65005.161	28	6	0.000
Test Length x Sample Size	3.457×10^{-6}	6.020	56	13.8	0.000
Test Length x b Dif.	2.179×10^{-3}	0.948	56	13.8	0.584
Test Length x Distribution	1.272×10^{-6}	7.856	56	13.8	0.000
Sample Size x b Dif.	7.060×10^{-9}	30.554	56	13.8	0.000
Sample Size x Distribution	2.486×10^{-11}	131.364	56	13.8	0.000
b Dif. x distribution	7.774×10^{-12}	177.200	56	13.8	0.000
Test Length x Sample Size x b Dif.	2.803×10^{-6}	1.514	112	32.7	0.087
Test Length x Sample Size x distribution	8.490×10^{-8}	2.681	112	32.7	0.001
Test Length x b Dif. x Distribution	3.189×10^{-6}	1.481	112	32.7	0.099
Sample Size x b Dif. x Distribution	2.672×10^{-11}	9.095	112	32.7	0.000

2. Multiple comparisons of power and Type I error rates of LR, MMLR7, MMLR9, MMLR10, HGLM15, HGLM16 and HGLM17. When repeated measure analysis of variance (Wilks' Lambda test) is used, for power, $\Lambda = 0.09$, F=124.28, degrees of freedom are 6 and 75, and $p < 0.0001$; for Type I error rate, $\Lambda = 0.17$, F=59.27, degrees of freedom are 6 and 75, and $p < 0.0001$. When the pairs that are of interest are

compared, the results of the paired t-tests are shown in the follow table. In every cell, the upper number is the average difference and the lower is its standard error, and * means $p < 0.0056$ by a Bonferroni correction.

		HGLM 15	HGLM 16	HGLM 17	MMLR 7	MMLR 9	MMLR 10
Power	RLR	0.10* 0.03	0.000 0.000	-0.006 0.003	0.059 0.030	-0.003* 0.0004	-0.009* 0.002
	HGLM 15				-0.04* 0.002		
	HGLM 16					-0.003* 0.0004	
	HGLM 17						-0.003* 0.001
Type I Error rate	RLR	-0.173* 0.030	0.000 0.000	-0.009* 0.002	-0.21* 0.030	-0.002* 0.0004	-0.008* 0.002
	HGLM 15				-0.04* 0.002		
	HGLM 16					-0.002* 0.0004	
	HGLM 17						-0.001 0.0004

3. Similarity rates of DIF detection between the 7 models with the IRT ability estimate.

Model	HGLM15	HGLM16	HGLM17	MMLR7	MMLR9	MMLR10
LR	70.30	100.00	94.81	67.98	99.73	94.83
HGLM15		70.30	70.40	95.94	70.22	70.47
HGLM16			94.82	67.99	99.73	94.84
HGLM17				68.09	94.80	99.70
MMLR7					67.94	68.16
MMLR9						94.83

REFERENCES

References

Agresti, A. (1997). A model for repeated measurements of a multivariate binary response. *Journal of the American Statistical Association*, 92, 315-321.

Bolt, D. M., & Gierl, M. J. (2006). Testing features of graphical DIF: application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement*, 43(4), 313-333

Breslow, N., & Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

Camilli, G. (2006). Test Fairness. In R. L. Brennan (Eds.), *Educational Measurement* (pp. 221-256). Westport, CT: Greenwood.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items: an NCME instructional module. *Educational Measurement: Issues and Practice*, 17(1), 31-44.

Creswell, M. (1991). A multilevel bivariate model. In R. Prosser, J. Rasbash, and H. Goldstein (Eds.). *Data Analysis with ML3*. London, Institute of Education.

Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (1983). *Assessing Unexpected Differential Item Performance of Female Candidates on SAT and TSWE Forms Administered in December 1977: An Application of the Standardization Approach* (RR-83-9). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23(4), 355-68.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, 33(4), 465-484.

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: a comparison of four methods. *Educational and Psychological Measurement*, 67(4), 565-582.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, 13(3), 378-402.

Giesbrecht, F. G., & Gumpertz, M. L. (2004). *Planning, Construction, and Statistical Analysis of Comparative Experiments.* Hoboken, NJ: Wiley & Sons.

Glonek, G., & McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society*, Series B, 57(3), 533-546.

Goldstein, H. (1995). *Multilevel Statistical Models*. London: Arnold.

Goldstein, H., & McDonald, R. (1988). A general model for the analysis of multilevel data. *Psychometrika*, 53(4), 455-467.

Goldstein, H., & Rasbash, J. (1996). Improved approximations to multilevel models with binary responses. *Journal of the Royal Statistical Society*, Series A, 159, 505-513.

Griffiths, P., Brown, J., & Smith, P. (2004). A comparison of univariate and multivariate multilevel models for repeated measures of use of antenatal care in Uttar Pradesh. *Journal of the Royal Statistical Society*, Series A, 167(4), 73-89.

Guo, G., & Zhao, H. (2000). Multilevel modeling for binary data, *Annual Review of Sociology*, 26, 441-62.

Hidalgo, M. D., & López-Pina, J. A. (2004) Differential item functioning detection and effect size: a comparison between logistic regression and Mantel-Haenszel Procedures, *Educational and Psychological Measurement*, 64(6), 903-915.

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

Hox, J. (2002). *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Erlbaum.

Humphreys, L. G., & Taber, T. (1973). Ability factors as a function of advantage and disadvantaged groups. *Journal of Educational Measurement*, 10(2), 107-115.

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.

Kamata, A. (1998). *Some Generalizations of the Rasch Model: an Application of the Hierarchical Generalized Linear Model.* Unpublished doctoral dissertation, Michigan State University.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement,* 38(1), 79–93.

Kamata, A., & Binici, S. (2003) *Random-effect DIF Analysis via Hierarchical Generalized Linear Model.* Paper presented at the biannual meeting of Psychometric Society, July 2003, Sardinia, Italy.

Kamata, A., Chaimongkol, S., Genc, E., & Bilir, K. (2005). *Random-Effect Differential Item Functioning Across Group Unites by the Hierarchical Generalized Linear Model.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Kim, W. (2003). *Development of a Differential Item Functioning Procedure Using the Hierarchical Generalized Linear Model: A Comparison Study with Logistic Regression Prodedure.* Unpublished doctoral dissertation, Pennsylvania State University.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, 65(6), 935-953.

Longford, N. (1993). *Random Coefficient Models.* Oxford, England: Clarendon Press.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Luppescu, S. (2002). *DIF detection in HLM item analysis.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Mcleod, A. (2001). Multivariate multilevel models. In A. H. Leyland & H. Goldstein (Eds.), *Multilevel Modelling of Health Statistics.* West Sussex, England: John Wiley & Sons.

Mellenberg, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, 7, 105-106.

Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, 46(4), 407-419.

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics*, 10(2), 133-142.

Narayanan, P., & Swaninathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, 20(3), 257-274.

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

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

Raudenbush, S., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks: Sage, second edition.

Raudenbush, S., Rowan, B. & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. High-School Data. *Journal of Educational Statistics*, 16(4), 295-330.

Raudenbush, S., Yang, M., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, 9(1), 141-157.

Rizopoulos, D (2006). ltm: an R Package for Latent Variable Modeling and Item Response Theory Analyses, *Journal of Statistical Software*, 17(5).

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society*, Series A, 158, 73-89.

SAS Institute Inc. (2008). *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Schabenberger, O., & Pierce, F. J. (2002). *Contemporary Statistical Models for the Plant and Soil Sciences,* Boca Raton, FL: CRC Press.

Shen, L. (1999). *A Multilevel Assessment of Differential Item Functioning*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Snijders, T., & Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

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

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J. & Feather, C. (2002). Analysis of Differential Item Functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, 27(1), 53-75.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Detection of differential item functioning using the parameters of item response theory models. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp.129-145). Hillsdale, NJ: Lawrence Erlbaum.

Thum, Y. M. (1997). Hierarchical linear models for multivariate outcomes. *Journal of Educational and Behavioral Statistics*, 22(1), 77-108.

Van den Bergh, H., Kuhlemeier & Wijnstra, J. (1995). *Differential Item Functioning from a Multilevel Perspective*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Van den Noortgate, W. & Boeck, P. D. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, 30(4), 443-464.

Van den Noortgate, W., Boeck, P. D., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics,* 28(4), 369-386.

Whitmore, M. L., & Schumacker, R. E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Applied Psychological Measurement,* 59(6), 910-927.

Wolfinger, R., & O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48(3-4), 233–243.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item Response Theory. In R. L. Brennan (Eds.), *Educational Measurement* (pp. 111-153). Westport, CT: Greenwood.

Yang, M., Goldestein, H. & Heath, A. (2000). Multilevel models for repeated binary outcomes: attitudes and voting over the electoral cycle. *Journal of the Royal Statistical Society*, Series A, 163(1), 49-62.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.