2013-06-06

Complex versus Simple Modeling for Differential Item functioning: When the Intraclass Correlation Coefficient (ρ) of the Studied Item is Less Than the ρ of the Total Score

Ying Jin
University of Miami, wenshuohua@hotmail.com

UNIVERSITY OF MIAMI

COMPLEX VERSUS SIMPLE MODELING FOR DIFFERENTIAL ITEM FUNCTIONING (DIF) DETECTION: WHEN THE INTRACLASS CORRELATION COEFFICIENT (ρ) OF THE STUDIED ITEM IS LESS THAN THE ρ OF THE TOTAL SCORE

By

Ying Jin

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida
June 2013

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

COMPLEX VERSUS SIMPLE MODELING FOR DIFFERENTIAL ITEM FUNCTIONING (DIF) DETECTION: WHEN THE INTRACLASS CORRELATION COEFFICIENT (ρ) OF THE STUDIED ITEM IS LESS THAN THE ρ OF THE TOTAL SCORE

Ying Jin

Approved:

_____
Nicholas D. Myers, Ph.D.
Associate Professor of Educational and Psychological Studies

_____
M. Brian Blake, Ph.D.
Dean of the Graduate School

_____
Batya Elbaum, Ph.D.
Associate Professor of Teaching and Learning

_____
Jaime Maerten-Rivera, Ph.D.
Research Associate of Teaching and Learning

_____
Etiony Aldarondo, Ph.D.
Associate Dean for Research

_____
Soyeon Ahn, Ph.D.
Assistant Professor of Educational and Psychological Studies

JIN, YING (Ph.D., Educational and Psychological Studies)

Complex versus Simple Modeling for Differential Item Functioning (DIF) Detection: When the Intraclass Correlation Coefficient (ρ) of the Studied Item is Less than the ρ of the Total Score (June 2013)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Nicholas D. Myers.
No. of pages in text. (55)

Previous research has demonstrated that DIF methods that do not account for multilevel data structure could result in too frequent rejection of the null hypothesis (i.e., no DIF) when the intraclass correlation coefficient (ρ) of the studied item was the same as ρ of the total score. The current study extended previous research by comparing the performance of DIF methods when ρ of the studied item was less than ρ of the total score, a condition that may be observed with considerable frequency in practice. The performance of two frequently used simple DIF methods that do not account for multilevel data structure, the Mantel-Haenszel test (MH) and Logistic Regression (LR), was compared to a less frequently used complex DIF method that does account for multilevel data structure, Hierarchical Logistic Regression (HLR). HLR and LR performed equivalently in terms of significance tests under most generated conditions, and MH was conservative across all conditions. Effect size estimates of HLR, LR and MH were more accurate and consistent under the Rasch model than under the 2 parameter item response theory model. The results of the current study provide evidence to help researchers further understand the comparative performance between complex and simple modeling for DIF detection under multilevel data structure.

ACKNOWLEDGEMENTS

I would like to thank the support from my mom, dad, and my little sister. I couldn't go this far without your love and support. My husband, Zhuo, who has accompanied me during the time here at US, has given me all his support and love to make me successful, this is for you.

My deepest appreciation goes to my advisor, Nick, who has helped me in every way to make me a good scholar. I remembered every minute we sat down together going through word by word, line by line of my dissertation. I couldn't make it without your kindness, patience, and guidance. You will be my forever advisor and friend.

I would also like to thank all my committee members who have provided extremely helpful advice and comments to make this work improved.

TABLE OF CONTENTS

Page

LIST OF TABLES……v

LIST OF FIGURES……vi

LIST OF ACRONYMS……vii

Chapter

1 INTRODUCTION……1

2 LITERATURE REVIEW……6

3 HYPOTHESES……17

4 METHODS……18

5 RESULTS……22

6 CONCLUSIONS……29

7 DISCUSSIONS……32

REFERENCES……36

TABLES……40

FIGURES……47

APPENDIX: R SYNTAX……54

LIST OF TABLES

Page

Table 1. An Illustration of a Two-Way Contingency Table at Level k of the Matching Variable.....40

Table 2. Item Parameters Used to Generate Items with Uniform DIF…..……...41

Table 3. $\rho_{y|x}$ for each level of manipulated factor.....42

Table 4. Type I error rate (standard deviation) for each level of manipulated factors.....43

Table 5. Power (standard deviation) for each level of manipulated factors.....44

Table 6. MSE across levels of J, n, and item type under the Rasch model.....45

Table 7. MSE across levels of J, n, and item type under the 2PL model.....46

LIST OF FIGURES

Page

Figure 1. Type I error rate for HLR, LR, and MH for Rasch and 2PL model at different levels of number of clusters when the grouping variable was at the cluster level (between).....47

Figure 2. Type I error rate for HLR, LR, and MH for Rasch and 2PL model at different levels of sample size with each cluster when the grouping variable was at the cluster level (between).....48

Figure 3. Type I error rate for HLR, LR, and MH for Rasch and 2PL model at different levels of item type when the grouping variable was at the cluster level (between).....49

Figure 4. Power for HLR, LR, and MH for Rasch and 2PL model at different levels of number of clusters when the grouping variable was at the cluster level (between).....50

Figure 5. Power for HLR, LR, and MH for Rasch and 2PL model at different levels of sample size with each cluster when the grouping variable was at the cluster level (between).....51

Figure 6. Power for HLR, LR, and MH for Rasch and 2PL model at different levels of item type when the grouping variable was at the cluster level (between).....52

Figure 7. Mean biases across levels of DIF methods and item type under the 2PL model 53

LIST OF ACRONYMS

DIF……...…..Differential Item Functioning

ES…….....….....Effect Size

HLR…….....…Hierarchical Logistic Regression

IRT…….....….....Item Response Theory

LR…….....…..... Logistic Regression

MH……...…..... Mantel-Haenszel Test

Chapter 1: Introduction

Differential item functioning (DIF) methods have been used to test for measurement invariance for decades (Osterlind & Everson, 2009). Analysis of measurement invariance is important because the independent relationship between group membership and the probability of correct responses after conditioning on the latent target ability is a necessary condition for item validity (Ackerman, 1992; Millsap & Meredith, 1992). When a dependent relationship exists, item validity can be jeopardized due to the presence of secondary abilities that an item is not intended to measure but that differ between groups (Camilli & Shepard, 1994). DIF methods test for the existence of such secondary abilities.

It is well known that applying standard statistical tests with multilevel data can result in inflated type I error rate (i.e., the probability of reject null hypothesis when the null hypothesis is correct) due to violation of the assumption of the independence of observations (Raudenbush & Bryk, 2002). In DIF analysis, the complex hierarchical logistic regression (HLR) method has recently been employed to account for multilevel data structure and its performance was compared to the simple logistic regression (LR) method (French & Finch, 2010). French and Finch (2013) further extended their research by examining the frequently used Mantel-Haenszel test (MH) under multilevel data structure. The results of these studies by French and Finch indicated that HLR effectively controlled type I error rate, whereas LR and MH resulted in more excessive false positives than the nominal level as the intraclass correlation coefficient (ρ) increased when the grouping variable was at the cluster level (e.g., school type). The current study extended the literature by (a) comparing HLR, LR, and MH with respect to significance

tests under a condition that may frequently be observed in practice, but that was not simulated in the French and Finch (2010; 2013) studies, and (b) comparing effect size (ES) estimates of the three methods examined in the French and Finch studies. The results of the current study, therefore, were expected to build on previous research by comparing the performance of simple versus complex DIF methods when ρ of the studied item was less than ρ of the total score, a condition that may be observed with considerable frequency in practice but has yet to be studied.

French and Finch (2010; 2013) were the first researchers who investigated the comparative performance between multiple DIF methods under multilevel data structure by simulation studies. The results and implications of French and Finch's studies, however, were based on the condition that the dichotomous multilevel data were generated where ρ of the studied item (ρ_y) was equal to ρ of the total score (ρ_x). In practice, a frequently observed situation is $\rho_y < \rho_x$ for dichotomous items (MacCallum, Zhang, Preacher & Rucker, 2002). Under multilevel data structure, both the studied item and the total score are governed by the underlying distribution of the latent ability with certain level ρ . The dichotomous studied item can be regarded as the dichotomization of the continuous latent ability by the item threshold. The total score, on the other hand, can be regarded as the continuous measure of the latent ability. Under classical test theory, high reliability of the total score can be achieved if the total score is a sufficient statistic for estimating the latent ability (e.g., under the Rasch model and with no DIF contamination, Baker & Kim, 2004), which will in turn result in more accurate estimation of the latent ability, leading to ρ_x being a more accurate estimate of ρ of the latent ability. On the contrary, MacCallum et.al. have shown that when the continuous latent

trait was dichotomized, the reliability of the dichotomized variable decreased. The decreased reliability can negatively bias estimation of ρ_y due to attenuated correlations.

Under multilevel data structure, ρ can be interpreted as the correlation between individuals within the same cluster (Bock, 1989). The estimated correlation is as follows (e.g., Crocker & Algina, 1986; Murphy & Davidshofer, 1988; Zimmerman & Williams, 1977):

$$\hat{\rho} = \hat{R}_{y_j,y_k} = R_{y_j,y_k}\sqrt{r_{y_j} r_{y_k}}\ , \qquad (1)$$

where r_{y_j} and r_{y_k} are the reliabilities of the scores (i.e., the studied item score or the total score), $\hat{R}_{y_j,y_k}$ (i.e., $\hat{\rho}$) and R_{y_j,y_k} (i.e., ρ) are the estimated and true correlations between person *j* and *k* within the same cluster, respectively. According to MacCallum et.al. (2002), r_{y_j} and r_{y_k} for the total score are greater than r_{y_j} and r_{y_k} for the dichotomous studied item because the dichotomization of the continuous latent trait can result in a decrease in reliability. With decreased reliability of the dichotomous studied item, $\hat{R}_{y_j,y_k}$ for the studied item will be smaller than $\hat{R}_{y_j,y_k}$ for the total score, therefore, $\rho_y < \rho_x$. Further evidence was provided by Caille, Leyrat, and Giraudeau's (2012) simulation study, where they compared ρ of the dichotomous and continuous outcome variables. The results indicated that ρ for the dichotomous outcome variable was consistently smaller than ρ for the continuous outcome variable.

The importance of the current study can be summarized in two aspects. First, under the condition of $\rho_y < \rho_x$, understanding the comparative performance between HLR, LR, and MH with respect to significance tests would help researchers evaluate the cost of

extra complexity of applying HLR in both theoretical and empirical studies. The relatively recent and still accumulating research on HLR for DIF detection and its complex modeling demand might be the reasons for HLR being a less frequently used DIF method than other regularly used DIF methods in practice. The simpler LR method (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990), and the industry standard Mantel-Haenszel test (MH test, Holland & Thayer, 1988, Penfield, 2007), on the other hand, have been intensively studied and compared (DeMars, 2009; French & Maller, 2007; Hidalgo & LOPez-Pina, 2004; Jodoin & Gierl; Narayanan & Swaminathan, 1996), but not under the condition of $\rho_y < \rho_x$. Determining whether or not these two simple DIF methods, LR and MH, could perform as well as HLR would provide evidence (a) to help researchers make the choice between the simple versus complex modeling for DIF detection, and (b) to question the necessity of applying HLR over LR and the MH test under the frequently observed condition of $\rho_y < \rho_x$ in practice.

Second, ES estimates have been used regularly to assist significance tests for the purpose of item retention or rejection (French & Maller, 2007; Hidalgo & LOPez-Pina, 2004; Jodoin & Gierl, 2001). No study, however, has compared HLR and other DIF methods with respect to ES estimates. Unlike standard errors, point estimates are relatively consistent and robust against the violation of the independence assumption (Raudenbush & Bryk, 2002). ES estimates based on point estimates of parameters are expected to be not severely affected by the dependency between observations as significance tests based on standard errors. Therefore, the ES estimates for the complex DIF method (i.e., HLR) accounting for multilevel data structure should be consistent with the ES estimates for the simple DIF method (i.e., LR) without accounting for multilevel

data structure. Different from HLR and LR, the MH test is a nonparametric DIF method and its ES estimate does not rely on parameter estimates. However, the MH common log odds ratio, which is used to describe the magnitude of DIF, is theoretically equivalent to the regression coefficient of the grouping variable in LR (DeMars, 2011; Swaminathan & Rogers, 1990). The ES estimate of the MH test, therefore, was expected to perform similarly to the ES estimate of LR. From an applied perspective, if ES estimates of the simple and computationally convenient DIF methods can accurately estimate the true DIF size, simple methods may be favored in some cases due to model parsimony.

Chapter 2: Literature Review

The following sections will review how HLR, LR, and MH methods are typically parameterized for DIF detection and how each method has been examined as a method for DIF detection in previous studies. A comparison of HLR and LR standard errors will also be provided to facilitate the hypotheses proposal in the current study.

Hierarchical Logistic Regression

HLR is generally used to account for multilevel data structure for dichotomous responses. The general model is written as

$$\begin{aligned} \eta_{ij} &= g(P(Y_{ij}=1 \mid X_{qij}, \mathrm{W}_{sj})) = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \ldots + \beta_{qj}X_{qij} \\ \beta_{qj} &= \gamma_{q0} + \gamma_{q1}W_{1j} + \gamma_{q2}W_{2j} + \ldots + \gamma_{qs}W_{sj} + \mu_{qj}, \end{aligned} \tag{2}$$

where Y_{ij} is the dichotomous item response for person *i* in cluster *j*, with 1 being correct, and 0 otherwise; *g* is the logit link; X_{qij} is the q^{th} person level predicator and W_{sj} is the s^{th} cluster level predictor; β and γ are the associated regression coefficients for *X* and *W*, respectively.

In DIF analysis, a person level covariate (i.e., usually the total score) is included to account for the between group mean ability difference in order to identify the true difference caused by DIF-present items. The total score is commonly used as the covariate and is denoted by *X*. In addition, a grouping variable is also included in the model for DIF detection. This grouping variable can be a within cluster grouping variable (e.g., G_{ij} =gender) or can be a between cluster grouping variable (e.g., G_j =school type). In order to utilize HLR for DIF detection, the covariate and the grouping variable are included in the general HLR model. When the grouping variable is at the individual level, Equation 2 is reduced to

$$\begin{aligned}
\eta_{ij} &= g(P(Y_{ij}=1 \mid X_{ij}, G_{ij})) = \beta_{0j} + \beta_{1j} X_{ij} + \beta_{2j} G_{ij} \\
\beta_{0j} &= \gamma_{00} + \mu_{0j} \\
\beta_{1j} &= \gamma_{10} \\
\beta_{2j} &= \gamma_{20}
\end{aligned} \tag{3}$$

When the grouping variable is at the cluster level, Equation 2 is reduced to

$$\begin{aligned}
\eta_{ij} &= g(P(Y_{ij}=1 \mid X_{ij}, G_{j})) = \beta_{0j} + \beta_{1j} X_{ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01} G_{j} + \mu_{0j} \\
\beta_{1j} &= \gamma_{10}
\end{aligned} \tag{4}$$

When the regression coefficients β_{2j} and γ_{01} in Equation 3 and 4 are significant, the studied item is identified to have uniform DIF (i.e., group difference is consistent across all levels of the latent trait) and therefore either should be removed from the test or sent to a panel of experts for revisions. The regression coefficients β_{2j} and γ_{01} can also be used as the uniform DIF ES estimates (Williams & Beretvas, 2006). Uniform DIF is a type of DIF that the difference between groups is consistent and unidirectional across all levels of the latent ability. Another type of DIF (i.e., nonuniform DIF: group difference is not consistent across all levels of the latent trait) can be detected by including an interaction term of the grouping variable and the covariate, but was not included because it was not of primary interest in the current study. From this point forward, DIF is referred as uniform DIF for textual parsimony.

In addition to French and Finch (2010) who examined the performance of HLR for DIF detection under multilevel data structure, HLR has also been used as a multilevel technique to explore potential causes for DIF. Previous studies have added either person characteristics (e.g., gender) or item characteristics (e.g., item word counts) as predictors to the second level to explain the different probabilities of correct responses after

conditioning on the latent ability (Kamata, 1999; Luppescu, 2000; Swanson, Clauser, Case, Nungster, & Featherman, 2002; Williams & Beretvas, 2006). In such studies, data were not multilevel in the sense that persons were nested within clusters, but were assumed to be independently collected. HLR was utilized in such studies where either items were nested within persons with person characteristics as the second level predictors, or persons were nested within items with item characteristics as the second level predictors. Three level models have been established to examine the relationship between latent traits and predictors at different levels, where items were nested within persons, and persons were nested within clusters (Pastor, 2003). No study, however, has utilized such three level models for exploring potential causes for DIF as well as accounting for multilevel data structure in DIF detection.

Logistic Regression

Similar to HLR, LR detects DIF relying on significance tests of the regression coefficients (Swaminathan & Roger, 1990). LR model for uniform DIF detection is written as

$$\eta_i = g(P(Y_i = 1 \mid X_i, G_i)) = \beta_0 + \beta_1 X_i + \beta_2 G_i \quad (5)$$

where Y_i is the dichotomous item response for person *i*. The probability of correct response for the studied item is modeled by the grouping variable G_i conditioning on the covariate X_i. The purified total score (i.e., computed from DIF-free items) is generally used as the covariate to remove the impact caused by group mean ability difference as in HLR. When β_2 is significant, the studied item has uniform DIF. The uniform DIF ES can also be estimated by β_2 (Penfield, 2007). As compared to HLR, LR is considered as the

simpler DIF method because it does not require complex modeling for DIF detection and generally requires smaller sample size than HLR.

A number of studies have investigated the performance of LR under various conditions (e.g., DeMars, 2009; Finch & French, 2007; Hidalgo & LOPez-Pina, 2004; Rogers & Swaminathan, 1993). The most noticeable advantage of LR is that it is more powerful in detecting nonuniform DIF than other DIF methods due to its model configuration that the LR model can include an interaction term of the grouping variable and the covariate for nonuniform DIF detection. However, the performance of LR for nonuniform DIF detection was not of primary interest of the current study, and is not discussed further in the following sections.

Standard Errors of Hierarchical Logistic Regression and Logistic Regression Coefficients

In order to further understand the comparative performance between HLR and LR under multilevel data structure in the current study, standard errors of HLR and LR should be examined first. It has been demonstrated that the relationship between ρ_x and ρ_y can influence the magnitude of standard errors of HLR regression coefficients through ρ of the residuals ($\rho_{y|x}$, Raudenbush, 1997). In the context of DIF analysis, $\rho_{y|x}$ is ρ of the residuals after removing the effect of the covariate (i.e., the total score) from the studied item. Results of previous studies have indicated that the magnitude of $\rho_{y|x}$ can be affected by the magnitude of ρ of the covariate (i.e., ρ_x): holding ρ_y constant, $\rho_{y|x}$ becomes smaller as ρ_x increased (Bloom, 2005; Hedges & Hedberg, 2007). In the current study, under the condition of $\rho_y < \rho_x$, $\rho_{y|x}$ was expected to decrease, and

therefore, to affect the comparative performance of HLR and LR in terms of standard errors of their regression coefficients.

According to Raudenbush (1997), when the grouping variable (*G*) is at the cluster level (G_j), the combined model can be written as

$$\eta_{ij} = \gamma_{00} + \gamma_{01} G_j + \gamma_{20} X_{ij} + \mu_{0j} \tag{6}$$

where $\eta_{ij} = g(P(Y_{ij} = 1 \mid X_{ij}, G_j))$ for person *i* in cluster *j*, and *g* is the logit link, G_j =1 for focal group and G_j =0 for reference group, X_{ij} is the person level covariate (i.e., the total score), and the random components $\mu_j \sim N(0, \tau^2_{y|x})$. The regression coefficient γ_{01} is used to determine the presence of DIF in terms of significance test (Williams & Beretvas, 2006). Raudenbush has derived that:

$$\begin{aligned}
Var(\hat{\gamma}_{01} \mid X)_{HLR} &= \frac{n\tau^2_{y|x} + \sigma^2_{y|x}}{Jnp(1-p)} \left(1 + \frac{U}{V\varphi_{HLR} + Y} \right), \\
\Delta_x &= \tau^2_x + \sigma^2_x / n, \\
U &= Jp(1-p)(M_{.F} - M_{.R})^2 / \Delta_x, \\
V &= SS_{wx} / \sigma^2_x, \\
Y &= SS_{bx} / \Delta_x, \\
\varphi_{HLR} &= \frac{(n\rho_{y|x} + 1 - \rho_{y|x})(1-\rho_x)}{(n\rho_x + 1 - \rho_x)(1-\rho_{y|x})}, \\
\rho_{y|x} &= \frac{\tau^2_{y|x}}{\tau^2_{y|x} + \sigma^2_{y|x}}, \\
\rho_x &= \frac{\tau^2_x}{\tau^2_x + \sigma^2_x},
\end{aligned} \tag{7}$$

where *p* is the proportion of clusters assigning to the reference group, *J* is the number of clusters, *n* is the number of persons within each cluster and is the same for all clusters, and $M_{.F}$ and $M_{.R}$ are the group means for the focal and reference groups, respectively. Standard errors derived by Raudenbush (1997) were not for binary data specifically. The standard errors of HLR regression coefficients for binary data, however, should follow

the same form as in Raudenbush with a different form of within-cluster variance (Spybrook, Bloom, Congdon, Hill, Martines & Raudenbush, 2011, p. 151-153).

When G is the grouping variable at the individual level (G_{ij}) with the assumption of equal group proportion within each cluster, equation (2) becomes

$$\eta_{ij} = \gamma_{00} + \gamma_{10} G_{ij} + \gamma_{20} X_{ij} + \mu_{0j} \tag{8}$$

and equation (3) becomes

$$Var(\hat{\gamma}_{10} \mid X)_{HLR} = \frac{\tau^2_{y|x} + \sigma^2_{y|x}}{Jnp(1-p)} \left(1 + \frac{U}{V\varphi_{HLR} + Y} \right). \tag{9}$$

In general, $Var(\hat{\gamma}_{01} \mid X)_{HLR}$ at the cluster level is greater than $Var(\hat{\gamma}_{10} \mid X)_{HLR}$ at the individual level, leading to higher type I error rate when the grouping variable is at the individual level (Bloom, 2005). French and Finch (2010) have shown that the type I error rate at the individual level was well controlled whereas HLR was more conservative when the grouping variable was at the cluster level.

When item responses for a particular item are modeled by LR ignoring multilevel data structure, equation (2) and (4) are reduced to

$$Y_i = \gamma_0 + \gamma_1 G_i + \gamma_2 X_i + e_i, \tag{10}$$

where the cluster index j is removed. The regression coefficient γ_1 is used to determine the presence of DIF in terms of significance test (Penfield, 2007). The variance of γ_1 then becomes

$$Var(\hat{\gamma}_1 \mid X)_{LR} = \frac{\sigma^2_{y|x}}{Jnp(1-p)} \left(1 + \frac{U}{V\varphi_{LR} + Y} \right), \tag{11}$$

$$\varphi_{LR} = \frac{1-\rho_x}{n\rho_x + 1 - \rho_x},$$

In order to compare equations (7), (9), and (11), φ_{HLR} and φ_{LR} need to be rewritten in the comparable forms as follows

$$\begin{aligned}
\varphi_{HLR} &= \frac{(n\rho_{y|x}+1-\rho_{y|x})(1-\rho_x)}{(n\rho_x+1-\rho_x)(1-\rho_{y|x})} \\
&= \frac{\dfrac{(n\rho_{y|x}+1-\rho_{y|x})}{(1-\rho_{y|x})}}{\dfrac{(n\rho_x+1-\rho_x)}{(1-\rho_x)}} \\
&= \frac{\dfrac{n\rho_{y|x}}{(1-\rho_{y|x})}+1}{\dfrac{n\rho_x}{(1-\rho_x)}+1} \\
&= \frac{\dfrac{\rho_{y|x}}{(1-\rho_{y|x})}+\dfrac{1}{n}}{\dfrac{\rho_x}{(1-\rho_x)}+\dfrac{1}{n}} \\
&= \frac{\dfrac{1}{\dfrac{1}{\rho_{y|x}}-1}+\dfrac{1}{n}}{\dfrac{1}{\dfrac{1}{\rho_x}-1}+\dfrac{1}{n}}
\end{aligned}$$

and

$$\begin{aligned}
\varphi_{LR} &= \frac{1-\rho_x}{n\rho_x+1-\rho_x} \\
&= \frac{1}{\dfrac{n\rho_x}{1-\rho_x}+1} \\
&= \frac{\dfrac{1}{n}}{\dfrac{\rho_x}{1-\rho_x}+\dfrac{1}{n}} \\
&= \frac{\dfrac{1}{n}}{\dfrac{1}{\dfrac{1}{\rho_x}-1}+\dfrac{1}{n}}
\end{aligned}$$

Differences between equations (7), (9), and (11) lie in two places: (a) Numerators $n\tau^2_{y|x}+\sigma^2_{y|x}$ vs. $\tau^2_{y|x}+\sigma^2_{y|x}$ vs. $\sigma^2_{y|x}$ in the ratio outside of the parenthesis, and (b) φ_{HLR} and φ_{LR}.

From (a) and (b), the differences between $Var(\hat{\gamma}_{10}\mid X)_{HLR}$, $Var(\hat{\gamma}_{01}\mid X)_{HLR}$, and $Var(\hat{\gamma}_{1}\mid X)_{LR}$ are due to the magnitude of $\rho_{y|x}$. When $\rho_{y|x}$ is close to 0, $\tau^2_{y|x}$ becomes smaller, leading to approximate equivalence of components in (a). When $\rho_{y|x}$ is close to 0, the numerator $\frac{1}{\frac{1}{\rho_{y|x}}-1}+\frac{1}{n}$ in φ_{HLR} approaches to the numerator $\frac{1}{n}$ in φ_{LR} in (b). Therefore, $Var(\hat{\gamma}_{10}\mid X)_{HLR}$, $Var(\hat{\gamma}_{01}\mid X)_{HLR}$, and $Var(\hat{\gamma}_{1}\mid X)_{LR}$ will become closer when $\rho_{y|x}$ is close to 0. In other words, when the covariate X (i.e., the total score) explains all or most of the between group variance, which leads to negligible $\rho_{y|x}$, the HLR model will be reduced to the LR model. Previous studies have provided evidence that when appropriate person level covariates were included in the model, a significant proportion of between cluster variance could be explained, resulting in reduced, sometimes, negligible $\rho_{y|x}$ (Hedges & Hedberg, 2007; Porter & Raudenbush, 1987).

Because item responses are conditional on the latent ability, when the total score can accurately estimate the latent ability, it is an appropriate person level covariate to account for the between cluster variance. In French and Finch (2010; 2013) studies, where data were generated from two-parameter Item Response Theory (IRT) model, the purified total score was used as the person level covariate and was not a sufficient statistic for estimating the latent ability due to the varying discrimination parameters across items

(Baker & Kim, 2004). Other than the effect of dependency between observations, the inflated type I error rate of LR may be partially caused by the fact that the covariate (i.e., purified total score) could not explain the between cluster variance as much as if the covariate was a sufficient statistic for latent ability estimation (e.g., purified total score under the Rasch model). Thus, when the covariate is an accurate estimate of the latent ability, and can explain most of the between-cluster variance (i.e., negligible $\rho_{y|x}$), HLR and LR will be fairly equivalent in DIF detection with respect to significance tests under multilevel data structure. The relatively demanding HLR may be not necessary for the purpose of accounting for multilevel data structure if the simple LR can produce equivalent standard errors as HLR does.

Mantel-Haenszel test

Different from LR and HLR, MH is a nonparametric approach because no model parameters are used for DIF detection. As illustrated in Table 1, the observed proportions of item responses are obtained to form *m* strata of two-way contingency tables, where *m* is the number of levels of the matching variable (i.e., the total score). MH significance test and its ES estimate are given by

$$MH-\chi^2=\frac{[\sum_{k=1}^{m}R_{rk}-\sum_{k=1}^{m}E(R_{rk})|-0.5]^2}{\sum_{k=1}^{m}Var(R_{rk})}, \qquad (12)$$

and

$$\lambda_{MH}=\ln(\frac{\sum_{k=1}^{m}\frac{R_{rk}W_{fk}}{N_{tk}}}{\sum_{k=1}^{m}\frac{R_{fk}W_{rk}}{N_{tk}}}). \qquad (13)$$

The elements in equations 12 and 13 are explained in Table 1. The chi-square test is computed by evaluating the squared discrepancy between the observed frequency and the predicted frequency. The DIF effect size estimate is the weighted average of log odds ratios across all levels of the matching variable. For the MH test, the utilization of the total score to create distinct strata is theoretically the same as including the covariate in the LR method, where the probability of correct responses are modeled after conditioning on/controlling for the effect of the covariate (DeMars, 2011; Swaminathan & Rogers, 1990). The performance of the MH test is therefore expected to perform equivalently as the LR method in DIF detection.

The MH test is popular in practice, due in part, to its computational convenience. Few studies have been conducted to investigate the performance of MH under multilevel data structure in educational and psychological studies. Previous studies conducted in other areas (e.g., medicine) have shown that MH can overly reject the null hypothesis (e.g., no treatment effect) due to the violation of the independence assumption when the primary sampling unit was at the cluster level (Donald & Donner, 1987; Rao & Scott, 1992; Weerasekera & Bennett, 1992). These studies, however, either did not include any covariate or include covariates providing information (e.g., demographic information) that was not as informative as the total score employed in MH for DIF analysis.

French and Finch (2013) examined the performance of MH for DIF analysis with multilevel data. Several MH adjustments (i.e., adjusting for multilevel data structure) were also compared with the standa rd MH with respect to significance tests. The results indicated that standard MH exhibited inflated type I error rate as ρ increased, and did not perform as effectively as some of the MH adjustments. The results and implications of

this study, however, were based on a condition also imposed in French and Finch (2010): $\rho_y = \rho_x$. Due to its popularity in empirical settings, the investigation of MH should be conducted and compared with HLR and LR to gain further understanding of the comparative performance between these DIF methods under the condition of $\rho_y < \rho_x$.

Chapter 3: Hypotheses

The following hypotheses regarding to the performance of the three DIF methods, HLR, LR, and MH under the frequently observed condition of $\rho_y < \rho_x$ were as follows.

Hypothesis 1. HLR, LR, and MH will perform comparably with respect to significance tests when $\rho_{y|x}$ is close to zero.

Hypothesis 2. The ES estimates of HLR, LR, and MH will be comparable under multilevel data structure.

Chapter 4: Methods

Six factors were included in the current Monte Carlo study to investigate the comparative performance of HLR, LR, and MH under the condition of $\rho_y < \rho_x$. Levels of model type (2 levels), ρ (5 levels), item type (3 or 4 levels depending on the IRT model), grouping variable (2 levels), number of clusters (3 levels), and sample size within each cluster (3 levels) were crossed to create 630 conditions. Each condition was replicated 1000 times.

Data Generation

Twenty dichotomous items were generated using the R package for statistical computing (R Core Team, 2010). Item parameters were obtained from Narayanan and Swaminathan (1996), and were used to generate item responses based on latent ability values generated from a standard normal distribution. Test length was not manipulated in the current study because previous studies have provided evidence that test length generally did not have as meaningful effect on uniform DIF detection as other factors (Rogers & Swaminathan, 1993; Swaminathan & Ragers, 1990); and twenty-item tests were not uncommon in previous studies (Donoghue & Allen, 1993; French & Finch, 2010; Williams & Beretvas, 2006).

Factors manipulated

Model type. Two types of IRT models were used to generate item responses: the Rasch model and the two-parameter IRT model (2PL). The purified total score (i.e., score computed from DIF-free items) was used as the covariate in the LR and HLR methods; and the total score including the studied item was used as the matching variable for the MH test (Zwick, 1990). The total score is a sufficient statistic for the estimation of the

latent ability under the Rasch model but not under the 2PL model (Baker & Kim, 2004). Whether or not the sufficiency of the total score can affect $\rho_{y|x}$ is important to understand the comparative performance of the three DIF methods.

Intraclass Correlation Coefficient **.** Five levels of ρ were simulated to reflect various magnitude of correlations between individuals within each cluster, and were the same as in French and Finch (2010; 2013) studies. The five levels were 0.05, 0.15, 0.25, 0.35, and 0.45, which were within the range of frequently observed ρ in empirical studies (Hedges, 2007; Hedges & Hedberg, 2007).

Item Type**.** Previous studies have demonstrated that item parameters have a significant effect on the performance of DIF methods, especially those methods relying on parameter estimates (DeMars, 2011; Donoghue & Allen, 1993; Rogers & Swaminathan, 1993). In the current study, different levels of item difficulty (*b*) and discrimination parameters (*a*) were combined to create 3 types of items showing uniform DIF for the Rasch model, and 4 types of items showing uniform DIF for the 2PL model. The item parameters of these types of items were illustrated in Table 2, which were the same as in Rogers and Swaminathan (1993) covering a wide and reasonable range of item parameters in practice. Not all levels of item difficulty and discrimination parameters were combined due to potential imprecise parameter estimation, leading to unreasonable item responses (e.g., high difficulty and low discrimination). One level of DIF size of 0.6 was selected to represent a meaningful and commonly observed DIF size as in previous studies (DeMars, 2009; Jodoin & Gierl, 2001). This moderate to large DIF size was generated for the purpose of computing power, estimating ES, and maintaining a

reasonable number of conditions within the current study. The DIF size was generated as the difference between the item difficulty parameters between groups.

Grouping variable. As in French and Finch (2010; 2013), two types of grouping variable were generated: between-cluster and within-cluster grouping variables. The between-cluster grouping variable was generated at the cluster level with 2 groups (e.g., classroom level characteristics); the within-cluster grouping variable was generated at the individual level with 2 groups (e.g., gender). For simplicity, only one grouping variable was included at each level for the current study, although more individual and cluster level characteristics can be included to identify potential sources of DIF (Swanson, Clauser, Case, Nungster, & Featherman, 2002; Williams & Beretvas, 2006).

Number of Clusters and Sample Size within Each Cluster. Three levels of number of clusters (J) were generated to reflect small (J=25), medium (J=50), and large (J=100) number of clusters. The performance of HLR in detecting uniform DIF with small number of clusters has not been investigated previously (French & Finch, 2010). Smaller number of clusters might cause HLR to yield biased parameter estimates, thus leading to less consistency in uniform DIF detection than LR and MH, which generally do not require large sample size. Three levels of sample size within each cluster (n) were generated to reflect small (n=10), medium (n=30), and large (n=50) number of individuals within each cluster.

Outcomes

The analyses of HLR and LR were conducted using Mplus 6.1 (Muthén & Muthén, 2010), and the analysis of MH was conducted using R. Type I error rate and power were computed under the generated conditions to evaluate hypothesis 1, where Type I error

rate was calculated as the ratio of number of DIF-free items falsely identified as DIF-present items out of all replications, and power was calculated as the ratio of number of DIF-present items correctly identified as DIF-present items out of all replications. Bias and mean square error (*MSE*) of ES estimates of the three DIF methods described in previous sections were computed as follows to evaluate hypothesis 2:

$$Bias(\hat{ES}) = E(\hat{ES}) - ES \tag{12}$$

$$MSE(\hat{ES}) = [Bias(\hat{ES})]^2 + Var(\hat{ES}), \tag{13}$$

where $\hat{ES}$ is the estimated effect size, and *ES* is the true DIF size. Mean bias and *MSE* were averaged across replications to evaluate precision and consistency of ES. Mean bias smaller than 0.05 were considered negligible based on the established criterion (Muthén & Muthén, 2002). Because the levels of the manipulated factors were not fully crossed due to different levels of the item type factor for the Rasch and 2PL models, two repeated measures ANOVA analyses were performed on bias and *MSE* for the Rasch and 2PL models, respectively. Significant main effects and interaction terms with partial effect size ($\hat{\eta}_p^2$) greater than 0.001 were discussed.

Chapter 5: Results

Magnitude of intraclass correlation coefficients of the residuals ($\rho_{y|x}$)

Before discussing the comparative performance between HLR, LR, and MH with respect to type I error rate, the magnitude of $\rho_{y|x}$ should be examined first to assist the inferences of the results. Table 3 provides $\rho_{y|x}$ for each level of the manipulated factors. Under both Rasch and 2PL models, $\rho_{y|x}$ increased as ρ increased; $\rho_{y|x}$ decreased as *J* and *n* increased; $\rho_{y|x}$ with grouping variable at the cluster level was smaller than when the grouping variable was at the within cluster level; $\rho_{y|x}$ for items with moderate *b* and low or high *a* was the smallest, and $\rho_{y|x}$ for items with low *b* and high *a* was the largest. In general, $\rho_{y|x}$ under the Rasch model was smaller than $\rho_{y|x}$ under the 2PL model across all levels of the manipulated factors, with $\rho_{y|x}$ ranged from 0.018 to 0.056 under the Rasch model, and from 0.023 to 0.076 under the 2PL model. None of $\rho_{y|x}$, however, was greater than 0.08.

Type I error rate

Table 4 provides type I error rate and its standard deviation for each level of the manipulated factors. Under the Rasch model, both HLR and LR controlled type I error rate reasonably well at the nominal alpha level (i.e., 0.05), with HLR being slightly conservative and LR being slightly liberal. Under the 2PL model, HLR effectively controlled type I error rate at the nominal alpha level. LR had type I error rate ranged from 0.049 to 0.064 across the conditions. In summary, HLR and LR performed equivalently (i.e., similar type I error rates) and consistently (i.e., similar standard

deviations of type I error rates) across the conditions under the Rasch model, with the average difference in type I error rate of 0.003 and in its standard deviation of 0.001. Under the 2PL model, HLR outperformed LR slightly, with the average difference in type I error rate of 0.005 and in its standard deviation of 0.006. The slight advantage of HLR over LR under the 2PL model can be further observed by the magnitude of $\rho_{y|x}$, where $\rho_{y|x}$ ranged from 0.023 to 0.076 under the 2PL model and 0.018 to 0.056 under the Rasch model. MH, in general, was found to be conservative across all conditions under both models with type I error rate ranged from 0.034 to 0.045.

As depicted by Figures 1-3, the comparative performance of HLR, LR, and MH was quite different for the two model types. Under the Rasch model, the comparative performance between the three methods was not affected by *J*, *n*, and item type. When ρ was smaller than 0.25, LR maintained type I error rate at the nominal level, with HLR being slightly conservative and MH being more conservative than HLR. Once ρ exceeded 0.25, HLR maintained type I error rate at the nominal level, with LR being slightly inflated and MH being conservative. The difference of the comparative performance among the three methods in terms of type I error rate was small. Under the 2PL model, however, the difference became larger as compared to that under the Rasch model. As ρ, *J*, and *n* increased, the effect of model type became more obvious, where the discrepancy between HLR and LR became larger, but type I error rate inflation of LR did not exceed 0.100. In general, HLR controlled type I error rate well; LR was more liberal than under the Rasch model, especially for items with moderate to high *b* and high *a*, and MH being conservative under most conditions except when ρ was 0.45 and n=50, where type I error rate of MH was 0.054.

Graphical comparisons between HLR, LR, and MH were not provided for the condition of grouping variable being at the individual level because the results were very similar to that of French and Finch (2010), where HLR and LR performed equivalently across levels of manipulated factors. MH, as in the condition of grouping variable being at the cluster level, was conservative across all conditions.

Power

Table 5 provides power and its standard deviation for simulation conditions. Power for HLR, LR, and MH were all above the acceptable level of 0.8. Under the Rasch model, LR outperformed HLR slightly, and MH was the least efficient DIF method in terms of both equivalency (i.e., power) and consistency (i.e., standard deviation of power). The average magnitude of differences between the three methods across all conditions, however, was as small as 0.007 for power and 0.014 for its standard deviation. As number of clusters and sample size within each cluster increased, power increased. Power decreased slightly as ρ increased for all three methods, but the magnitude of increase did not exceed 0.006 for each method. When the grouping variable was at the cluster level, power was slightly lower than when the grouping variable was at the individual level, the magnitude of differences across three methods, again, was negligible (i.e., 0.001 to 0.006). Finally, items with moderate *b* performed better in terms of both DIF detection rates and consistency than items with either low or high *b*. In summary, even with tiny differences between levels of factors, all three methods can detect DIF efficiently under the Rasch model because the minimum detection rate was above the acceptable level of 0.8 (i.e., 0.851) across all levels of manipulated factors.

Under the 2PL model, with slight decrease in power and increase in standard deviations under most conditions, similar patterns were observed as under the Rasch model for all three methods, where LR being the most powerful method, and MH being the least efficient method.

The minimum detection rate of the three methods across levels of manipulated factors was 0.848, which was slightly lower than that under the Rasch model, but was above the acceptable level of 0.80.

Figure 4-6 depicted the comparative performance of HLR, LR, and MH under different levels of *J*, n, and item type, respectively, across all levels of ρ. As depicted by Figures 4 and 5, the comparative performance of HLR, LR, and MH was affected the most by *J* and n. When *J* and n were medium to large, power of HLR, LR, and MH all increased rapidly to 1.00 regardless of the magnitude of ρ, and the three DIF methods performed equivalently. When *J* and n was small, the discrepancy between the three methods was larger than when *J* and n was medium to large, with LR being the most efficient method and MH being the least efficient method. Another noticeable effect under the condition of small *J* and n was that power of HLR, LR, and MH all decreased as ρ increased. As depicted by Figure 6, item type has no effect on power across levels of ρ because similar patterns were observed across all levels of item type: HLR, LR and MH were comparative in terms of correctly detecting DIF-present items.

Bias

Under the Rasch model, mean bias of HLR, LR, and MH ES estimates ranged from -0.04 to 0.03, -0.04 to 0.05, and 0.01 to 0.06, respectively, across all simulation conditions. ES estimates of HLR and LR accurately estimated the true DIF size because mean biases

across all conditions were lower than the cut-off value of negligible mean bias of 0.05. MH odds ratio was positively biased under the condition of low *b* and high *b* with small number of clusters and sample size within each cluster, where the bias was only 0.01 higher than the cut-off value. Although main effects of DIF method ($p < 0.01$, $\hat{\eta}_p^2 = 0.027$), grouping variable ($p < 0.01$, $\hat{\eta}_p^2 = 0.004$), sample size ($p < 0.01$, $\hat{\eta}_p^2 = 0.002$) and the two-way interaction of DIF method by grouping variable ($p < 0.01$, $\hat{\eta}_p^2 = 0.007$) were significant, the mean biases of different levels of the main effects and the crossed levels of the interaction effect were smaller than the cut-off value, and therefore were considered negligible.

Under the 2PL model, mean bias of HLR, LR, and MH ES estimates ranged from -0.25 to 0.13, -0.25 to 0.16, and -0.23 to 0.18, respectively, across all levels of manipulated factors. ES estimates of HLR, LR and MH did not seem to be effective in estimating DIF size because mean biases across most conditions were higher than the cut-off value of 0.05. As under the Rasch model, the main effects of DIF method ($p < 0.01$, $\hat{\eta}_p^2 = 0.027$), grouping variable ($p < 0.01$, $\hat{\eta}_p^2 = 0.004$), sample size within each cluster ($p < 0.01$, $\hat{\eta}_p^2 = 0.002$) and the two-way interaction of DIF method by grouping variable ($p < 0.01$, $\hat{\eta}_p^2 = 0.008$) were significant, but were not discussed further due to negligible mean biases across levels of main effects and interaction effects. The main effect of item type ($p < 0.01$, $\hat{\eta}_p^2 = 0.625$) and the two-way interaction of DIF method by item type ($p < 0.01$, $\hat{\eta}_p^2 = 0.005$), however, needed to be discussed because the inaccuracy of ES estimates of the three DIF methods were attributed to the effect of item type. As illustrated by Figure 7, mean biases of HLR, LR, and MH ES estimates were negatively

biased when items had low *a*, and were positively biased when items with high *a*. Also, as *b* increased, mean bias also increased. Among ES estimates of the three DIF methods, HLR ES estimate was the least biased whereas MH odds ratio was the most biased when the items had high *a*.

MSE

Under the Rasch model, *MSE* of HLR, LR, and MH ES estimates ranged from 0.004 to 0.154, 0.004 to 0.162, and 0.004 to 0.205, respectively. Under the 2PL model, *MSE* of HLR, LR, and MH ES estimates ranged from 0.007 to 0.201, 0.008 to 0.217, and 0.013 to 0.280, respectively. Under both Rasch and 2PL models, as illustrated in Table 6 and Table 7, ES estimates of HLR and LR performed equivalently in terms of consistency with negligible average difference in *MSE* (Rasch: 0.002; 2PL: 0.003). MH odd ratio was the least consistent ES estimate among the three DIF ES estimates. Main effects of DIF method (both Rasch and 2PL: $p < 0.01$, $\hat{\eta}_p^2 = 0.014$), item type (Rasch: $p < 0.01$, $\hat{\eta}_p^2 = 0.067$; 2PL: $p < 0.01$, $\hat{\eta}_p^2 = 0.090$), number of clusters (Rasch: $p < 0.01$, $\hat{\eta}_p^2 = 0.318$; 2PL: $p < 0.01$, $\hat{\eta}_p^2 = 0.309$), and sample size within each cluster (Rasch: $p < 0.01$, $\hat{\eta}_p^2 = 0.430$; 2PL: $p < 0.01$, $\hat{\eta}_p^2 = 0.289$) were significant. The highest-order interaction terms containing these four significant main effects were also significant (Rasch: $p < 0.01$, $\hat{\eta}_p^2 = 0.001$; 2PL: $p < 0.01$, $\hat{\eta}_p^2 = 0.002$) under both models. For both Rasch and 2PL models, as number of clusters and sample size within each cluster increased, *MSE* decreased accordingly. And when an item had either low or high *b*, *MSE* was larger as compared to an item with moderate *b* for ES estimates of all three DIF methods. Under the 2PL model, however, as observed in Table 5, when an item with moderate *b* also had

low *a*, *MSE* was highest among four levels of item type as number of cluster and sample size with each cluster increased.

Chapter 6: Conclusions

As emphasized by French and Finch (2010; 2013), understanding the performance of DIF methods under multilevel data structure is crucial to item and test validity. Results of DIF detection can be spurious if a DIF method is inconsistent with the data structure, leading to invalid item and test scores. From an applied perspective, misspecified DIF-present or DIF-free items may obscure systematic difference in test scores between groups of interest within- (e.g., gender) or between-clusters (e.g. schools). The current study extended French and Finch's studies by comparing DIF methods that may or may not comply with data structure under a condition that may be observed frequently in practice, but has not been studied. The comparative performance of a complex DIF method, HLR, and two simpler DIF methods, LR and MH, was examined by testing two hypotheses with respect to significance tests and ES estimates.

The first hypothesis of HLR, LR, and MH being comparable in terms of significance tests when $\rho_{y|x}$ was close to zero was tested under various conditions and was partially supported. When the grouping variable was at the cluster level, HLR, LR and MH performed equivalently in terms of controlling type I error rate at the nominal alpha level when ρ was small (i.e., $\rho < 0.25$) under both Rasch and 2PL models. When ρ became larger (i.e., $\rho > 0.25$), HLR generally outperformed LR with respect to type I error rate, and MH was slightly conservative under both models. The magnitude of type I error rate inflation of LR as compared to HLR and MH under the Rasch model, however, was not as large as that under the 2PL model, because the larger $\rho_{y|x}$ was, the more HLR outperformed LR and MH. When ρ was greater than 0.25, the average $\rho_{y|x}$ under the Rasch model was 0.034, and was 0.054 under the 2PL model, which confirmed the more

consistent and equivalent performance of the three DIF methods under the Rasch model than under the 2PL model. As compared to French and Finch (2010; 2013), the results of the current study were similar to theirs when the grouping variable was at the individual level, where all three methods performed equivalently in terms of maintaining type I error rate at the nominal level. The maximum magnitude of type I error rate inflation of LR under the 2PL model, however, was 0.440 under the condition of $\rho_y = \rho_x$ in French and Finch (2010), and was 0.167 under the condition of $\rho_y < \rho_x$ in the current study under similar conditions.

In regard to power, all three methods maintained power above the acceptable level (0.8) with trivial differences across all levels of manipulated factors in the current study. The effects of number of clusters and sample size within each cluster were most noticeable as compared to other factors (i.e., item type and intraclass correlation coefficient). The power of HLR, LR and MH were comparative and higher than the acceptable level when number of clusters and sample size within each cluster were medium to high, and was not affect by the magnitude of intraclass correlation coefficient. When number of clusters and sample size within each cluster was small, the relative advantage of LR over HLR and MH was more apparent than under the condition of medium to large number of clusters and sample size within each cluster, with power of HLR being slightly lower than power of LR, and power of MH being the lowest. Power of the three DIF methods, although was lower under the condition of small number of cluster and small sample size within each cluster than under the condition of medium to large number of clusters and sample size within each cluster, was above the acceptable level of 0.8. As compared to the results of French and Finch (2010; 2013), under the

similar conditions where DIF size was 0.6 in their studies, power of these three methods reached 0.8 only when the number of clusters and sample size within each cluster were large, which might be explained by the condition of $\rho_y = \rho_x$ generated in their studies.

The results of the current study partially supported the second hypothesis of HLR, LR, and MH being comparable in terms of ES estimates. Under the Rasch model, both HLR and LR ES estimates were equivalent in terms of estimating DIF size accurately under all conditions. MH ES estimate was also precise under most conditions except for either easy or difficulty items when the number of clusters and sample size within each cluster were small, where MH ES estimate was slightly positively biased. Under the 2PL model, the effect of item type was the most noticeable among all other factors. When items had low *a*, ES estimates of all three methods were negatively biased with negligible differences between them; when items had high *a* and either moderate or high *b*, ES estimates of all three methods were all positively biased with HLR ES estimate being the most precise, and MH ES estimate being the least precise. As for *MSE* of ES estimates, HLR, LR, and MH ES estimates were more consistent under the Rasch model than under the 2PL model with average difference as small as to the second decimal place. Within both Rasch and 2PL model, HLR and LR ES estimates performed equivalently in terms of consistency with average difference in the third decimal place, and MH was the least consistent DIF ES estimate with average difference in the second decimal place as compared to HLR and LR. In summary, the second hypothesis of all three ES estimates being comparable was supported under the Rasch model across all conditions, and under certain conditions of 2PL model (e.g., low *b* and low *a*).

Chapter 7: Discussions

For practitioners, the question brought up earlier regarding employing complex versus simple modeling for DIF detection with multilevel data under the condition of $\rho_y < \rho_x$ can be informed by the results of the current study. Under the Rasch model, the simple DIF method LR can be used as the most effective method when the number of cluster, sample size within each cluster, and ρ is small (i.e., smaller than 0.25), where HLR and MH can be slightly conservative with respect to type I error rate. Situations where ρ was greater than 0.25 is not as common as ρ was less than 0.25 based on empirical literature (What Works Clearinghouse, 2008). Therefore, complex DIF method might not be necessary under certain conditions, especially when researchers have limited resources to recruit participants in empirical settings. When ρ is greater than 0.25, the complex method HLR should be employed because LR can exhibit slightly inflated type I error rate and MH can be a little conservative. However, the consequences caused by slightly liberal LR and conservative MH with respect to significance tests can be compensated for by the precise and consistent ES estimates of LR and MH. Therefore, HLR, LR, and MH can be used interchangeably under the Rasch model, with simple DIF methods of LR and MH being more favorable when resources are limited or when model parsimony is of interest to researchers.

Under the 2PL model, researchers should use caution when simple DIF methods are employed with multilevel data structure being ignored. Under conditions of small ρ, small sample size within each cluster, LR can still be used interchangeably with HLR for easy and high discriminating items. When ρ becomes larger, LR is no longer appropriate. As discussed previously, because the purified total score under the 2PL model is not a

sufficient statistic for estimating the latent ability, including it as the personal level covariate can cause misleading results with respect to both significance tests and ES estimates. Unless a modified sufficient person level covariate under the 2PL model (e.g., $\sum aX_{ij}$, where a is the discrimination parameter, and X_{ij} is the person level score of person i in cluster j, Baker & Kim, 2004) is used in LR, HLR is generally preferred. On the other hand, MH, even under the 2PL model, can maintain reasonably high power to correctly identify DIF-present items but can be a little conservative in identifying DIF-free items. MH common odds ratio, however, is the least precise and stable ES estimate, which might be due to the same reason of insufficiency of the total score used as the matching variable. Therefore, MH ES estimate should be used with caution under the 2PL model.

Previously, regular DIF methods that didn't account for multilevel data structure (e.g., LR or MH) were employed with multilevel data structure ignored. These studies might be reexamined if the intraclass correlation coefficients were high and sample sizes were small with misspecified model (e.g., item responses followed the 2PL model were modeled by the Rasch model). Researchers should be cautious when using regular DIF methods (e.g., MH) with multilevel data. Based on the results of the current study, it is recommended that the magnitude of intraclass correlation coefficient and model-data fit should be examined first before selecting an appropriate DIF method, especially when number of clusters and sample size within each cluster were small.

A few factors that are sometimes included in DIF analysis were not manipulated in the current study, which may limit the generalizability of the results of the current study. For example, the results of the current study indicated that HLR, LR, and MH can

maintain reasonably high power across all conditions. This can be attributed to the fact that a medium to high level of DIF size (0.6) was generated. It would not be surprising to see that power would decrease if smaller DIF size was generated. However, as compared to the results of French and Finch (2010; 2013), power of the three DIF methods in the current study was higher under similar conditions, providing some evidence that the three DIF methods under the condition of $\rho_y < \rho_x$ in the current study might be more powerful than under the condition of $\rho_y = \rho_x$. In addition, DIF contamination, which had an unfavorable effect on several DIF methods (Finch, 2005; Woods, 2009) was not included in the current study, partly because the effect of DIF contamination can be alleviated by iterative item purification procedures or by including anchor items (Wang, Shih & Yang, 2009). Some other factors (e.g., test length) were not included based on the principle of avoiding confounding effects which can be somewhat removed by external remedies (e.g., inclusion of anchor items).

Results from the current study suggest that the literature in DIF analysis under multilevel data structure could be expanded in at least two ways. First, although LR and MH can effectively identify DIF-present items and retain DIF-free items under certain conditions, these methods were not consistent with multilevel data structure and therefore performed poorly under unfavorable conditions. A possible remedy may be to apply modifications to these DIF methods. For example, design effect (i.e., it inflates standard errors of parameter estimates to correct for the dependency between observations) can be accommodated in the associated significance tests to account for the multilevel data structure (Weerasekera & Bennett, 1992), and can be extended in DIF analysis for future research. Second, based on the results of previous and current studies, HLR is still

considered the most effective DIF method with respect to both significance tests and ES estimates under multilevel data structure, especially when the observations within each cluster are highly correlated. There are empirical situations where ρ can be high, and HLR should be used as the more appropriate method for DIF detection. For example, Myers, Feltz, Maier, Wolfe, and Reckase (2006) have demonstrated that ρ can be as high as 0.33 for physical education settings. In addition, HLR has great potential in DIF analysis and should be examined further because it was not only studied to examine its capability to account for multilevel data structure, but was also used to explore the potential causes of DIF (Kamata, 1999; Swanson, Clauser, Case, Nungster & Featherman, 2002; Williams & Beretvas, 2006). For example, three-level HLR models where items can be nested within persons (i.e., to identify potential DIF sources caused by item characteristics) and persons were nested within clusters (i.e., DIF detection that accounts for the dependency between observations) can be employed in DIF analysis for future studies. More research can be done under both conditions of $\rho_y < \rho_x$ and $\rho_y = \rho_x$ to obtain further understanding of complex versus simple modeling for DIF detection.

References

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

Baker, F.B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. NY: Marcel Dekker.

Bloom, H. S. (2005). *Learning more from social experiments: Evolving analytic approaches.* New York, NY: Russell Sage Foundation.

Bock, R. D. (1989). *Multilevel analysis of educational data*. CA: San Diego.

Caille, A., Leyrat, C., & Giraudeau, B. (2012). Dichotomizing a continuous outcome in cluster randomized trials: impact on power. *Statistics in Medicine, 31*, 2822-2832.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, Holt, TX: Rinehart and Winston.

DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, 34*(2), 149-170.

DeMars, C. E. (2011). An analytic comparison of effect sizes for differential item functioning. *Applied Measurement in Education, 24*, 189-209.

Donald. A., & Donner, A. (1987). Adjustments to the Mantel-Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. *Statistics in Medicine, 6*, 491-499.

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

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278-295.

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement, 47*, 299-317.

French, B. F., & Finch, W. H. (2013). Extensions of Mantel-Haenszel for Multilevel DIF Detection. *Educational and Psychological Measurement*. DOI: 10.1177/0013164412472341

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67*(3), 373-393.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*, 341-370.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Hidalgo, M. D., & LOPez-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 procedure. In Wainer, H. & Braun, H.I. Editor (Eds.), *Test validity* (pp.129-145). Erlbaum, Hillsdale, NJ.

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), 329-49.

Kamata, A. (1999). *Some generalizations of the Rasch model: An application of the hierarchical generalized linear model*. Unpublished doctoral dissertation, Michigan State University, East Lansing.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19-40.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92.

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

Murphy, K., & Davidshofer, C. (1988). *Psychological testing: Principles and applications*. Englewood Cliffs, NJ: Prentice Hall.

Muthén, L.K. and Muthén, B.O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4*, 599-620.

Muthén, L. K., & Muthén, B. O. (2010). Mplus: Statistical Analysis with Latent Variables (version 6.1) [Computer software]. Los Angeles, CA: Author.

Myers, N. D., Feltz, D. L., Maier, K. S., Wolfe, E. W., & Reckase, M. D. (2006). Athletes' evaluations of their head coach's coaching competency. *Research Quarterly for Exercise and Sport, 77*, 111–121.

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

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Los Angeles, CA: Sage.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal Of Educational Measurement, 44*(3), 187-210.

Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology, 34*(4), 383-293.

R Core Team. (2010). R: A Language and Environment for Statistical Computing (version 2.12.0) [Computer software]. Vienna, Austria: Author.

Rao, J. N. K., & Scott, A. J. (1992). A simple method for the analysis of clustered binary data. *Biometrics, 48*, 577-585.

Raudenbush, S. W. (1997). Statistical analysis and optical design for cluster randomized trials. *Psychological Methods, 2*(2), 173-185.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

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

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. W. (2011). Optimal Design for Longitudinal and Multilevel Research: documentation for the Optimal Design Software Version 3.0. Available from www.wtgrantfoundation.org.

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., Nungster, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics, 27*, 53-75.

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

Wang, W., Shih, C.-L., & Yang, C.-C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement, 69*, 713-731.

Weerasekera, D. R., & Bennett, S. (1992). Adjustments to the Mantel-Haenszel test for data from stratified multistage surveys. *Statistics in Medicine, 11*, 603-616.

What Works Clearinghouse. (2008). Procedures and standards handbook (version 3.0). Retrieved March 9, 2013, from http://ies.ed.gov/ncee/wwc/pdf/wwc_procedures_ v2_standards_ handbook.pdf.

Williams, N. J., & Beretvas, N. S. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement, 30*, 22-42.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*(1), 1-27.

Zimmerman, D. W., & Williams, R. H. (1977). The theory of test validity and correlated errors of measurement. *Journal of Educational Measurement, 19*, 125-134.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*(3), 185-193.

Table 1. *An illustration of a two-way contingency table at level k of the matching variable*

	Correct	Incorrect	Total
Focal	R_{fk}	W_{fk}	N_{fk}
Reference	R_{rk}	W_{rk}	N_{rk}
Total	R_{tk}	W_{tk}	N_{tk}

Note. b = Difficulty parameter; a = Discrimination parameter. R_{fk} = Number of correct responses for the focal group; R_{rk} = Number of correct responses for the reference group; R_{tk} = Number of correct responses for both focal and reference groups; W_{fk} = Number of incorrect responses for the focal group; W_{rk} = Number of incorrect responses for the reference group; W_{tk} = Number of incorrect responses for both focal and reference groups; N_{fk} = Total number of correct and incorrect responses for the focal group; N_{rk} = Total number of correct and incorrect responses for the focal group; N_{tk} = Total number of correct and incorrect responses for both focal and reference groups;

Table 2. *Item parameters used to generate items with uniform DIF*

Rasch model					2PL				
Item type	Reference		Focal		Item type	Reference		Focal	
	a	*b*	*a*	*b*		*a*	*b*	*a*	*b*
Low *b*	1.00	-1.81	1.00	-1.19	Low *b*, high *a*	1.20	-1.81	1.20	-1.19
Moderate *b*	1.00	-0.31	1.00	0.31	Moderate *b*, low *a*	0.60	-0.31	0.60	0.31
High *b*	1.00	1.19	1.00	1.81	Moderate *b*, high *a*	1.20	-0.31	1.20	0.31
					High *b*, high *a*	1.20	1.19	1.20	1.81

Note. *b* = Difficulty parameter; *a* = Discrimination parameter. 2PL: Two parameter item response model. Under the Rasch model, discrimination parameters are fixed to 1 across items, whereas discriminations parameters of the 2PL model can vary across items.

Table 3. $\rho_{y|x}$ *for each level of manipulated factors*

Factors and levels	Rasch	2PL
ρ		
0.05	0.029	0.030
0.15	0.030	0.036
0.25	0.032	0.043
0.35	0.034	0.052
0.45	0.038	0.063
Item type: Rasch (2PL)		
low b (high a)	0.041	0.076
moderate b (high a)	0.026	0.040
high b (high a)	0.032	0.041
moderate b (low a)		0.023
J		
25	0.039	0.049
50	0.033	0.045
100	0.027	0.041
n		
10	0.056	0.064
30	0.024	0.037
50	0.018	0.033
Grouping variable		
within	0.035	0.048
between	0.030	0.042

Note. ρ = Intraclass correlation coefficient; J = Number of clusters; n = Sample size within each cluster; 2PL: Two parameter item response model.

Table 4. *Type I error rate (standard deviation) for each level of manipulated factors*

Factors and levels	Rasch			2PL		
	HLR	LR	MH	HLR	LR	MH
ρ						
0.05	0.048(0.008)	0.050(0.008)	0.040(0.009)	0.047(0.006)	0.050(0.006)	0.039(0.007)
0.15	0.048(0.007)	0.051(0.007)	0.040(0.008)	0.050(0.007)	0.053(0.009)	0.041(0.008)
0.25	0.049(0.007)	0.053(0.009)	0.039(0.008)	0.049(0.007)	0.055(0.012)	0.040(0.008)
0.35	0.049(0.006)	0.054(0.009)	0.040(0.007)	0.051(0.008)	0.059(0.017)	0.040(0.009)
0.45	0.049(0.008)	0.053(0.011)	0.039(0.008)	0.052(0.008)	0.064(0.026)	0.043(0.010)
Item type: Rasch (2PL)						
low b (high a)	0.048(0.007)	0.052(0.011)	0.038(0.008)	0.050(0.007)	0.054(0.009)	0.043(0.008)
moderate b (high a)	0.048(0.007)	0.052(0.008)	0.041(0.008)	0.049(0.008)	0.060(0.026)	0.039(0.010)
high b (high a)	0.049(0.007)	0.053(0.007)	0.040(0.007)	0.050(0.007)	0.057(0.014)	0.041(0.008)
moderate b (low a)				0.049(0.007)	0.054(0.011)	0.040(0.008)
J						
25	0.047(0.007)	0.051(0.010)	0.036(0.008)	0.049(0.008)	0.055(0.016)	0.037(0.008)
50	0.050(0.007)	0.054(0.008)	0.041(0.007)	0.050(0.007)	0.056(0.016)	0.041(0.008)
100	0.048(0.007)	0.051(0.009)	0.042(0.007)	0.051(0.007)	0.058(0.016)	0.044(0.009)
n						
10	0.047(0.008)	0.049(0.007)	0.034(0.008)	0.048(0.007)	0.051(0.008)	0.035(0.008)
30	0.049(0.007)	0.053(0.008)	0.042(0.007)	0.051(0.007)	0.057(0.014)	0.042(0.007)
50	0.049(0.007)	0.054(0.010)	0.043(0.006)	0.051(0.007)	0.061(0.022)	0.045(0.008)
Grouping variable						
within	0.049(0.007)	0.049(0.007)	0.040(0.008)	0.050(0.007)	0.049(0.007)	0.040(0.008)
between	0.047(0.007)	0.055(0.009)	0.040(0.008)	0.050(0.008)	0.064(0.019)	0.041(0.009)

Note. ρ = Intraclass correlation coefficient; J = Number of clusters; n = Sample size within each cluster; HLR: Hierarchical logistic regression; LR: Logistic regression; MH: Mantel-Haenszel; 2PL: Two parameter item response model.

Table 5. *Power (standard deviation) for each level of manipulated factors*

	Rasch			2PL		
Factors and levels	HLR	LR	MH	HLR	LR	MH
ρ						
0.05	0.963(0.091)	0.965(0.089)	0.951(0.120)	0.962(0.096)	0.963(0.094)	0.950(0.124)
0.15	0.965(0.089)	0.966(0.087)	0.951(0.121)	0.960(0.098)	0.962(0.095)	0.949(0.125)
0.25	0.962(0.094)	0.964(0.091)	0.949(0.124)	0.960(0.099)	0.962(0.095)	0.948(0.124)
0.35	0.961(0.096)	0.963(0.092)	0.949(0.125)	0.956(0.103)	0.959(0.099)	0.945(0.129)
0.45	0.959(0.099)	0.960(0.096)	0.945(0.130)	0.954(0.107)	0.957(0.103)	0.943(0.133)
Item type: Rasch (2PL)						
low b (high a)	0.950(0.111)	0.951(0.109)	0.934(0.145)	0.927(0.142)	0.930(0.138)	0.914(0.166)
moderate b (high a)	0.984(0.043)	0.985(0.041)	0.977(0.061)	0.955(0.100)	0.957(0.096)	0.941(0.133)
high b (high a)	0.952(0.106)	0.954(0.102)	0.936(0.142)	0.990(0.028)	0.991(0.025)	0.986(0.040)
moderate b (low a)				0.962(0.087)	0.964(0.083)	0.947(0.122)
J						
25	0.905(0.142)	0.908(0.137)	0.873(0.187)	0.899(0.151)	0.904(0.146)	0.871(0.190)
50	0.982(0.032)	0.983(0.031)	0.975(0.045)	0.978(0.044)	0.979(0.042)	0.971(0.055)
100	1.000(0.001)	1.000(0.001)	0.999(0.002)	0.999(0.003)	0.999(0.003)	0.999(0.004)
n						
10	0.890(0.135)	0.893(0.131)	0.851(0.177)	0.883(0.146)	0.887(0.141)	0.848(0.182)
30	0.996(0.007)	0.997(0.006)	0.996(0.008)	0.994(0.014)	0.995(0.012)	0.993(0.015)
50	1.000(0.001)	1.000(0.000)	1.000(0.000)	0.999(0.002)	1.000(0.001)	1.000(0.001)
Grouping variable						
within	0.965(0.087)	0.964(0.089)	0.950(0.122)	0.962(0.094)	0.961(0.097)	0.947(0.126)
between	0.959(0.100)	0.963(0.091)	0.948(0.125)	0.955(0.107)	0.960(0.097)	0.946(0.127)

Note. ρ = Intraclass correlation coefficient; J = Number of clusters; n = Sample size within each cluster; HLR: Hierarchical logistic regression; LR: Logistic regression; MH: Mantel-Haenszel; 2PL: Two parameter item response model.

Table 6. *MSE across levels of J, n, and item type under the Rasch model*

J	*n*	Item type		Rasch	
25	10	low b	0.135	0.139	0.187
		moderate b	0.079	0.085	0.098
		high b	0.137	0.140	0.183
	30	low b	0.041	0.044	0.051
		moderate b	0.026	0.029	0.030
		high b	0.043	0.046	0.052
	50	low b	0.025	0.027	0.030
		moderate b	0.016	0.018	0.018
		high b	0.026	0.027	0.030
50	10	low b	0.062	0.065	0.080
		moderate b	0.039	0.043	0.047
		high b	0.067	0.069	0.081
	30	low b	0.021	0.022	0.025
		moderate b	0.013	0.015	0.015
		high b	0.021	0.023	0.025
	50	low b	0.013	0.014	0.015
		moderate b	0.008	0.009	0.009
		high b	0.013	0.014	0.015
100	10	low b	0.031	0.032	0.038
		moderate b	0.019	0.021	0.023
		high b	0.032	0.034	0.039
	30	low b	0.011	0.012	0.013
		moderate b	0.007	0.008	0.008
		high b	0.010	0.011	0.012
	50	low b	0.007	0.007	0.008
		moderate b	0.004	0.005	0.005
		high b	0.007	0.007	0.008

Note. b = Difficulty parameter; *a* = Discrimination parameter; *J* = Number of clusters; *n* = Sample size within each cluster.

Table 7. *MSE across levels of J, n, and item type under the 2PL model*

J	*n*	Item type	2PL		
25	10	low b, high a	0.160	0.165	0.230
		moderate b, high a	0.096	0.107	0.126
		high b, high a	0.189	0.197	0.268
		moderate b, low a	0.115	0.115	0.118
	30	low b, high a	0.050	0.055	0.067
		moderate b, high a	0.032	0.040	0.044
		high b, high a	0.058	0.065	0.078
		moderate b, low a	0.077	0.074	0.071
	50	low b, high a	0.032	0.035	0.042
		moderate b, high a	0.021	0.028	0.031
		high b, high a	0.036	0.042	0.049
		moderate b, low a	0.071	0.068	0.064
50	10	low b, high a	0.073	0.077	0.099
		moderate b, high a	0.047	0.055	0.063
		high b, high a	0.087	0.093	0.116
		moderate b, low a	0.085	0.084	0.082
	30	low b, high a	0.025	0.028	0.035
		moderate b, high a	0.018	0.024	0.028
		high b, high a	0.032	0.037	0.045
		moderate b, low a	0.068	0.065	0.061
	50	low b, high a	0.016	0.018	0.024
		moderate b, high a	0.012	0.018	0.021
		high b, high a	0.020	0.025	0.031
		moderate b, low a	0.065	0.061	0.057
100	10	low b, high a	0.036	0.039	0.050
		moderate b, high a	0.025	0.032	0.037
		high b, high a	0.044	0.049	0.060
		moderate b, low a	0.072	0.069	0.066
	30	low b, high a	0.013	0.015	0.021
		moderate b, high a	0.010	0.016	0.019
		high b, high a	0.017	0.022	0.028
		moderate b, low a	0.063	0.060	0.056
	50	low b, high a	0.008	0.010	0.015
		moderate b, high a	0.008	0.013	0.016
		high b, high a	0.013	0.017	0.022
		moderate b, low a	0.062	0.058	0.054

Note. b = Difficulty parameter; *a* = Discrimination parameter; *J* = Number of clusters; *n* = Sample size within each cluster; 2PL: Two parameter item response model.

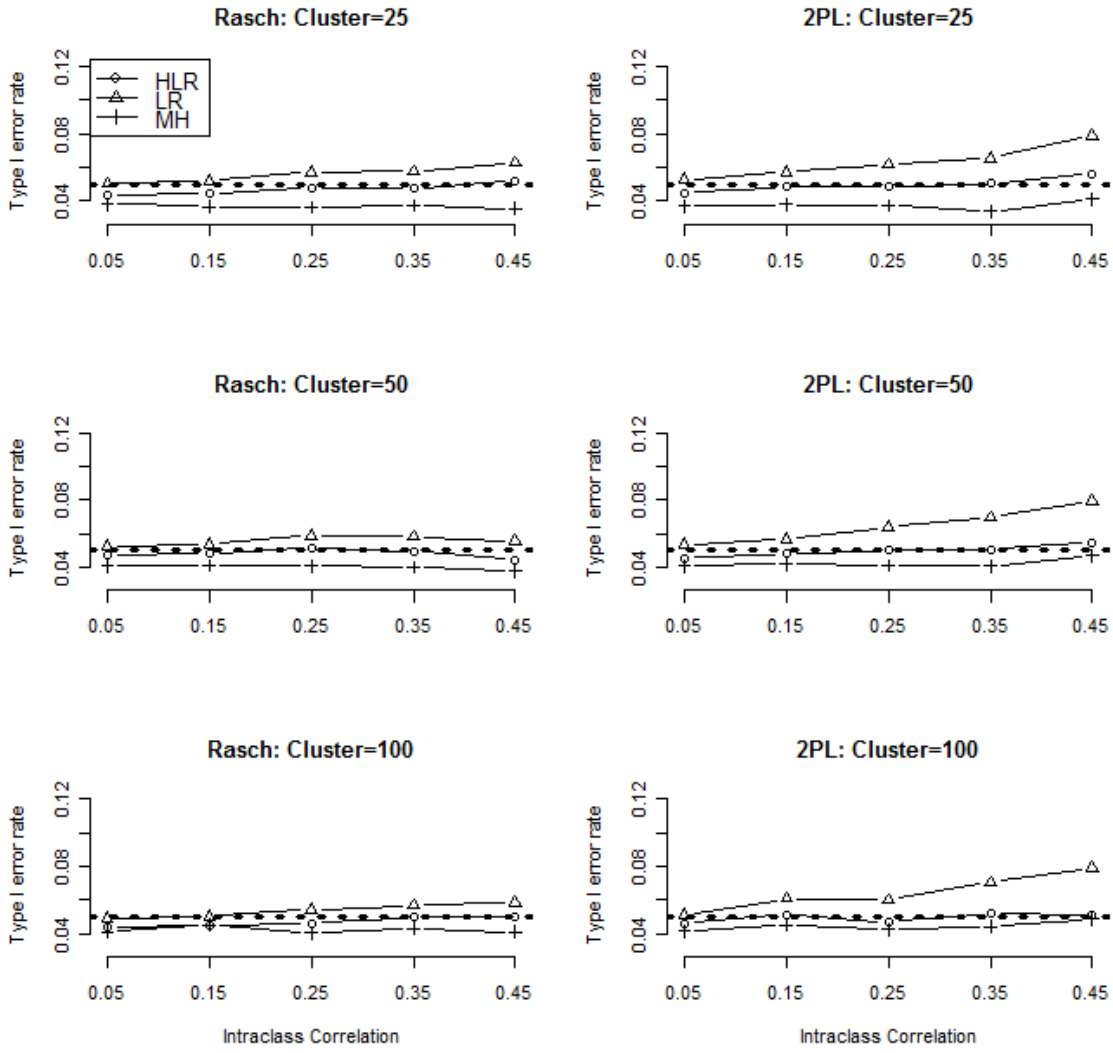

Figure 1. Type I error rate for HLR, LR, and MH for Rasch and 2PL model at different levels of number of clusters when the grouping variable was at the cluster level (between). The dotted line indicated the cut-off value of 0.05.

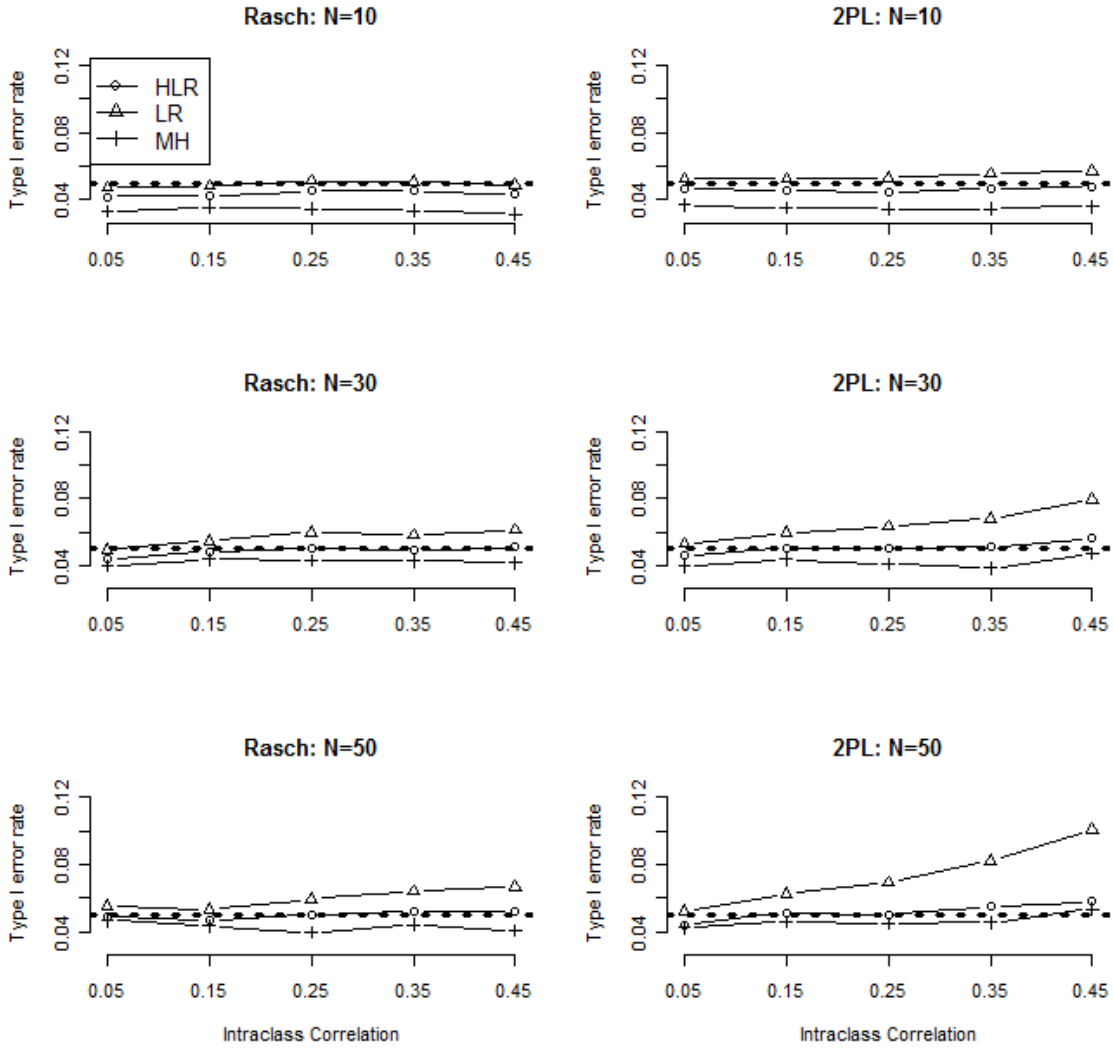

Figure 2. Type I error rate for HLR, LR, and MH for Rasch and 2PL model at different levels of sample size with each cluster when the grouping variable was at the cluster level (between). The dotted line indicated the cut-off value of 0.05.

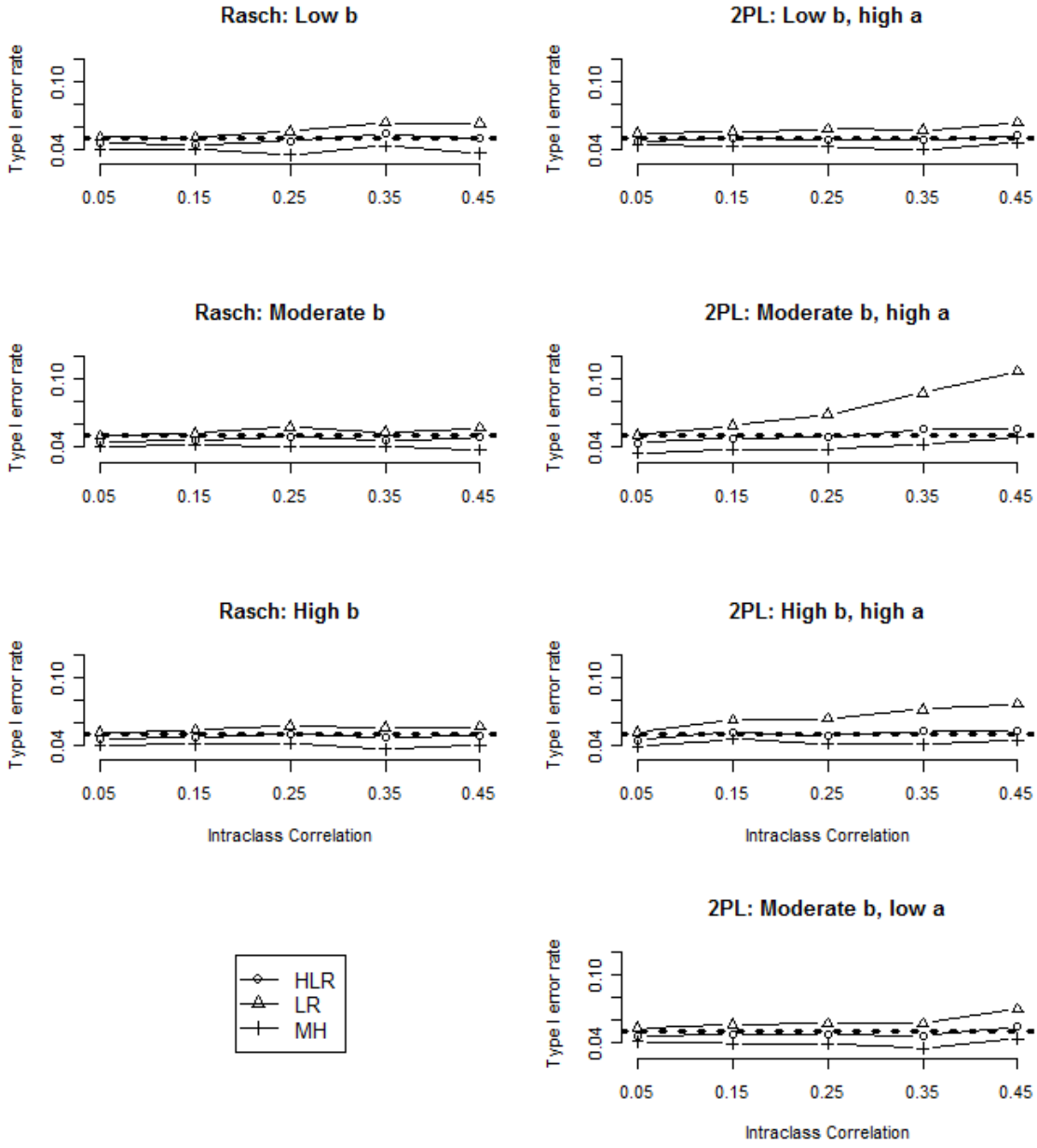

Figure 3. Type I error rate for HLR, LR, and MH for Rasch and 2PL model at different levels of item type when the grouping variable was at the cluster level (between). The dotted line indicated the cut-off value of 0.05.

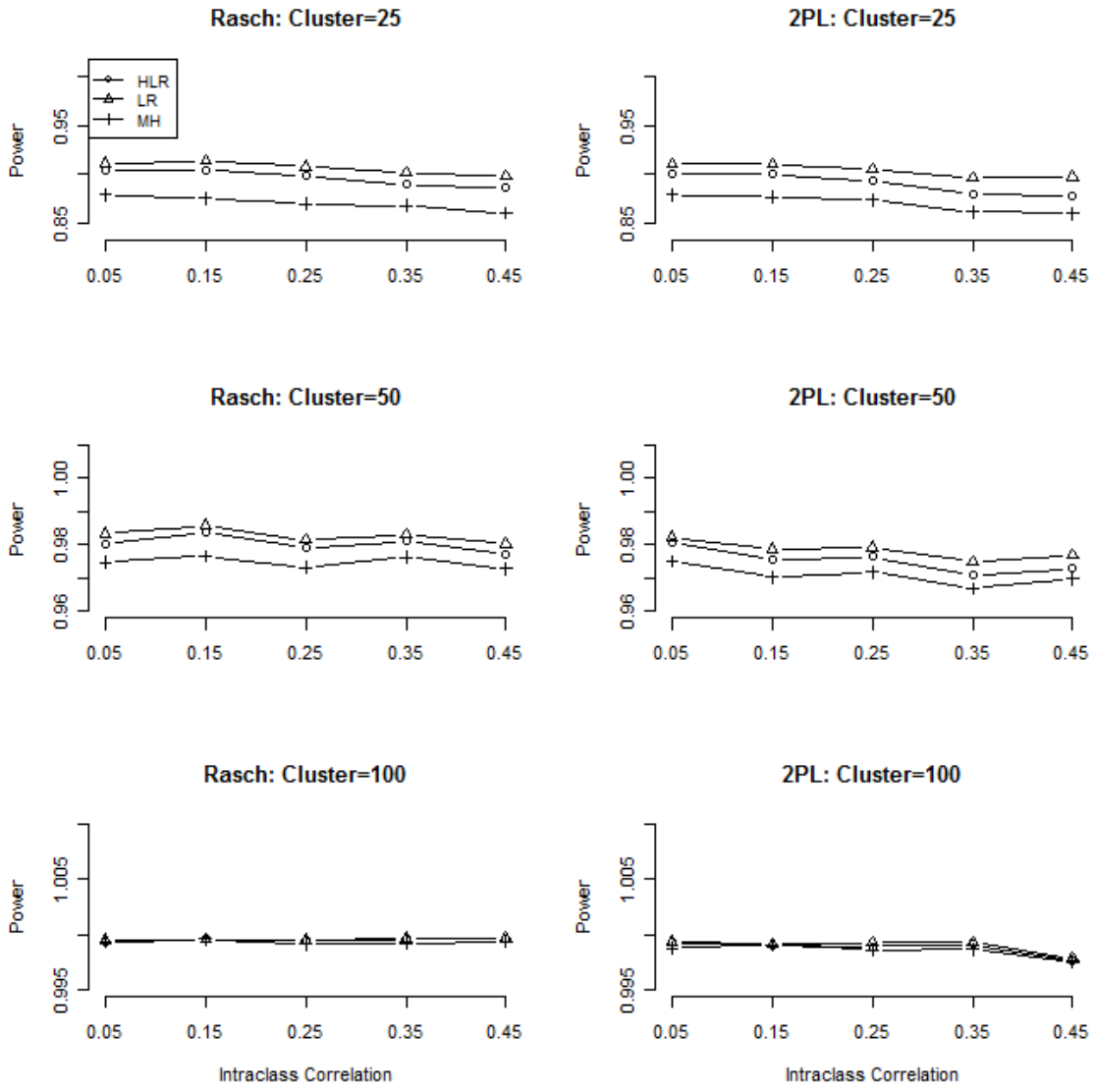

Figure 4. Power for HLR, LR, and MH for Rasch and 2PL model at different levels of number of clusters when the grouping variable was at the cluster level (between).

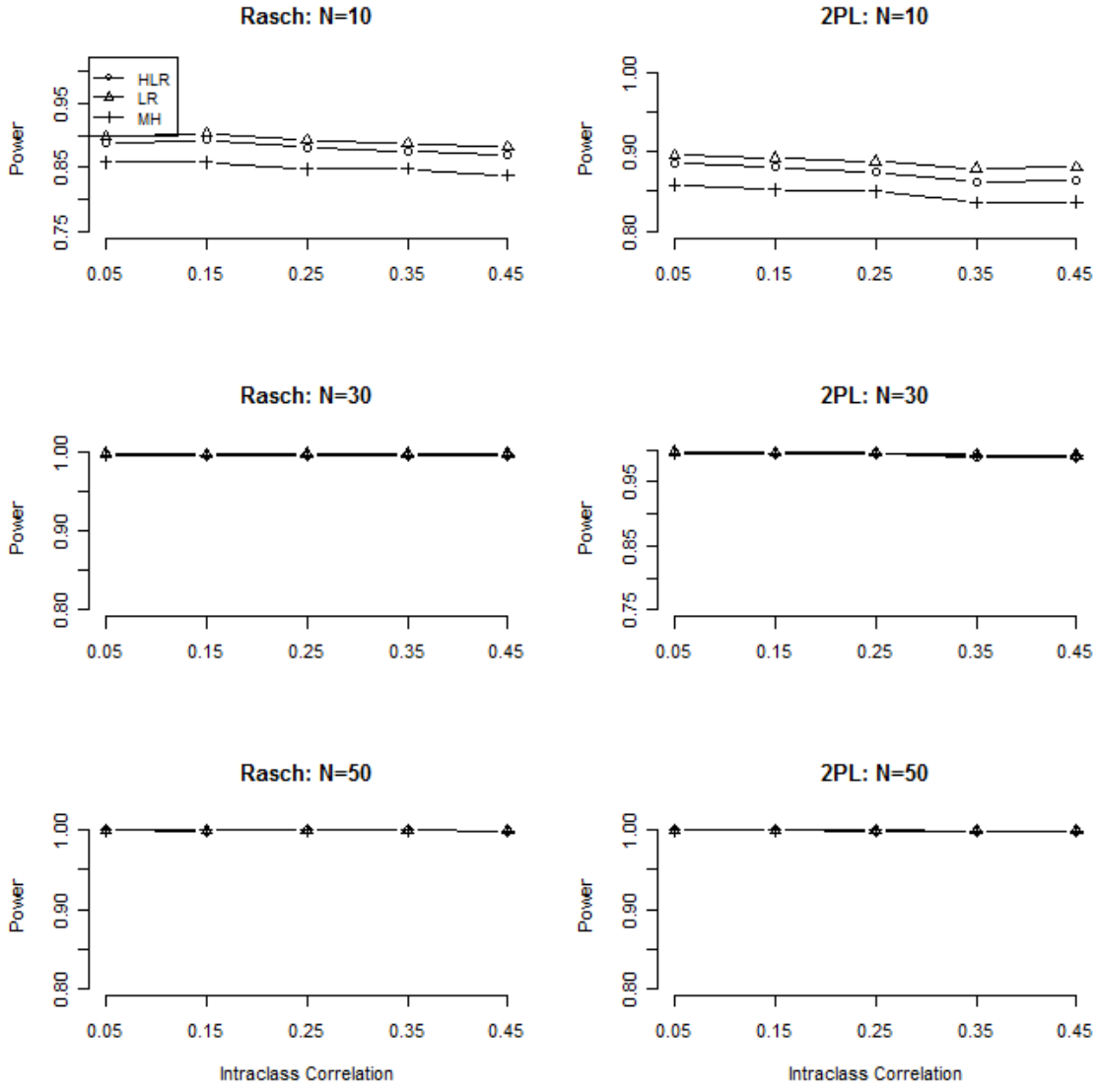

Figure 5. Power for HLR, LR, and MH for Rasch and 2PL model at different levels of sample size with each cluster when the grouping variable was at the cluster level (between).

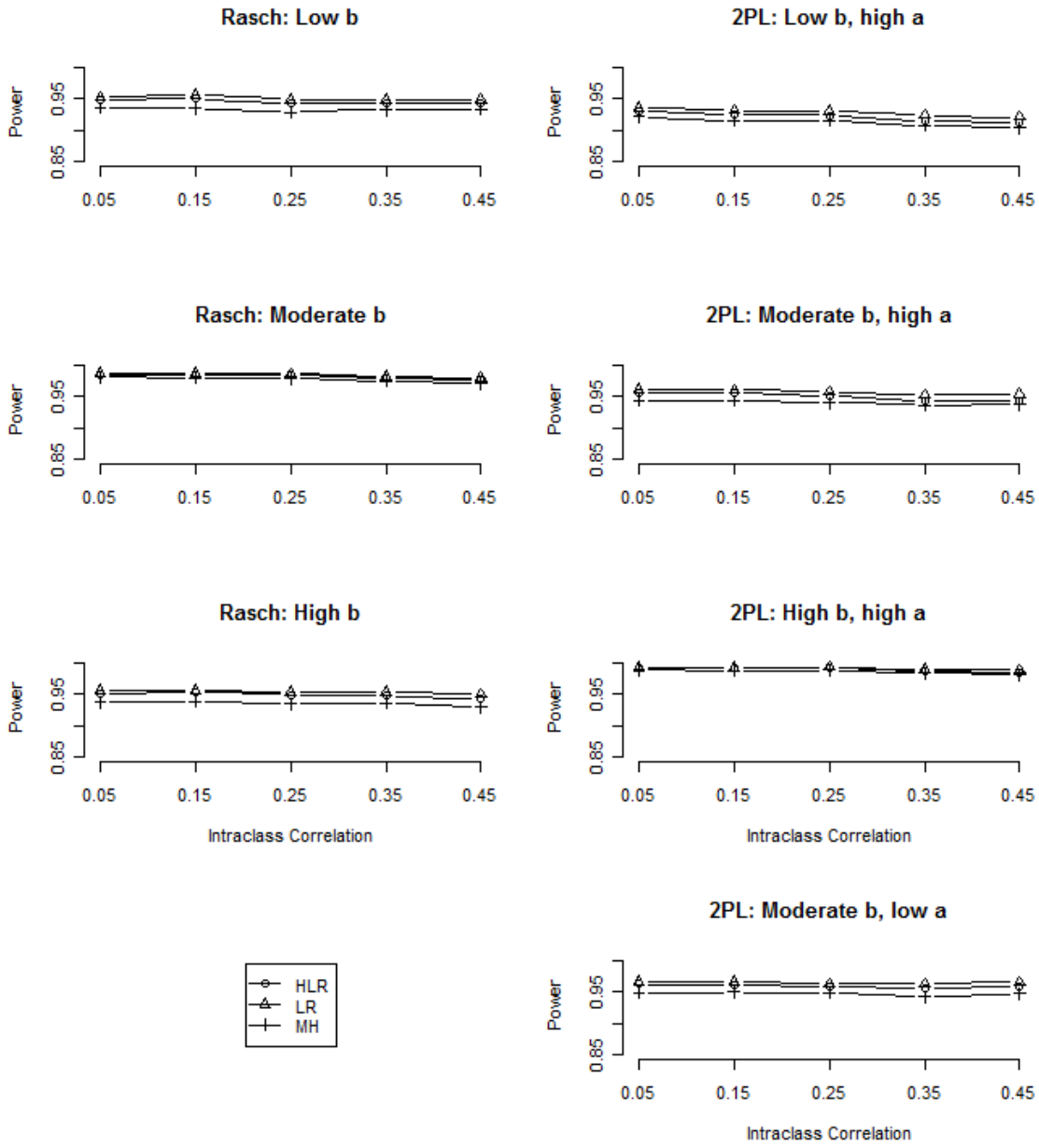

Figure 6. Power for HLR, LR, and MH for Rasch and 2PL model at different levels of item type when the grouping variable was at the cluster level (between).

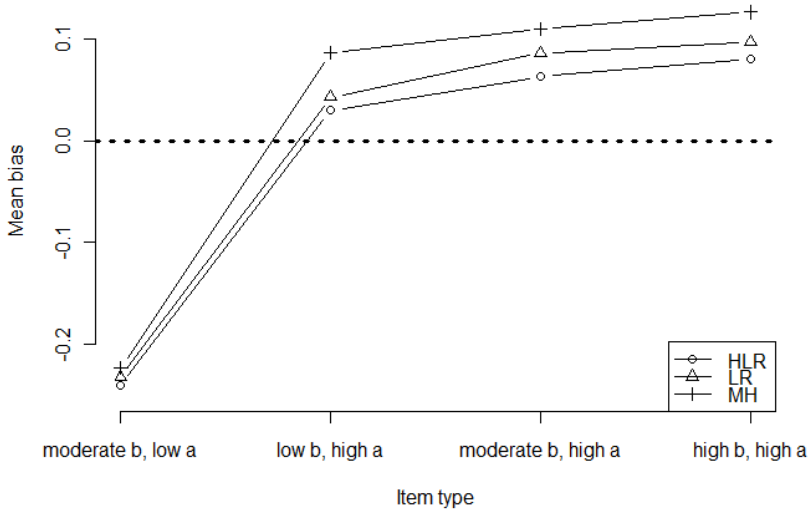

Figure 7. Mean biases across levels of DIF methods and item type under the 2PL model. The dotted line indicated the reference line of 0.00.

Appendix: R syntax

```
condition=function(k,n,m1,m2,c_sigma,c_tau,gamma00,parameter){
{
### k = level 2 units
### n = level 1 units
########## generating multilevel ability values ##########
all<-list()
db<-matrix(0,nrow=n,ncol=4,byrow=TRUE)
for (j in 1:k)
 {tau<-rnorm(1,0,sqrt(c_tau))
                                  for (i in 1:n)
                                    {sigma<-rnorm(1,0,sqrt(c_sigma))
                                     y<-gamma00+sigma+tau
                                     db[i,]<-cbind(i,sigma, tau, y)}
all[[j]]<-cbind(j,db)
                            }
library(gdata)
dataset<-combine(all[[1]])
for (j in 2:k){
dataset<-rbind(dataset,combine(all[[j]]))}
dataset<-dataset[1:5]
dataset=cbind(dataset,c(1:(n*k)))
colnames(dataset)<-c("l2id", "l1id", "sigma", "tau", "y", "sid")
}
############ creating within and between group variable ##############
### n = level 1 units
### m1 = # of reference group for within group values generation
### m2 = # of reference group for between group values generation
### 0 presents reference group, 1 represents focal group
### k = level 2 units
within.group=function(n,m1,k){
within=function(n,m1){
vec=c(1:n)
test=sample(1:n,m1)
for (b in 1:length(test)){
vec[test[b]]=0
vec2=ifelse(vec==0,0,1)}
return=vec2}
withinvalues=stack(data.frame(replicate(k,within(n,m1))))[,1]
return=withinvalues}

between.group=function(n,m2,k){
vec=c(1:k)
test=sample(1:k,m2)
for (b in 1:length(test)){
```

```
vec[test[b]]=0
vec2=ifelse(vec==0,0,1)}
mx=matrix(0,nrow=n,ncol=k)
for (d in 1:k){
mx[,d]=rep(vec2[d],n)
betweenvalues=stack(data.frame(mx))[,1]}
return=betweenvalues}

info.data=cbind(dataset[,1:2],within.group(n,m1,k),between.group(n,m2,k),dataset[,5])
colnames(info.data)=c("l2id","l1id","within.v","between.v","theta")

############ function for generating response matrix ##############
    pix <- function(iparam, thp){
                                    (exp(1.7*iparam[,2]*
                                    (thp-iparam[,1]))/(1+exp(1.7*iparam[,2]*(thp-
iparam[,1]))))
                            }

person.within.ref=subset(info.data,info.data[,3]==0)
person.within.foc=subset(info.data,info.data[,3]==1)
person.between.ref=subset(info.data,info.data[,4]==0)
person.between.foc=subset(info.data,info.data[,4]==1)

out.within.ref=matrix(0,nrow=length(person.within.ref[,1]),ncol=20)
out.within.foc=matrix(0,nrow=length(person.within.foc[,1]),ncol=20)
out.between.ref=matrix(0,nrow=length(person.between.ref[,1]),ncol=20)
out.between.foc=matrix(0,nrow=length(person.between.foc[,1]),ncol=20)

for(e in 1:20){
out.within.ref[,e]=ifelse(pix(parameter[e,1:2],person.within.ref[,5]) >=
runif(length(person.within.ref[,1])),1,0)
out.within.foc[,e]=ifelse(pix(parameter[e,3:4],person.within.foc[,5]) >=
runif(length(person.within.foc[,1])),1,0)
out.between.ref[,e]=ifelse(pix(parameter[e,1:2],person.between.ref[,5]) >=
runif(length(person.between.ref[,1])),1,0)
out.between.foc[,e]=ifelse(pix(parameter[e,3:4],person.between.foc[,5]) >=
runif(length(person.between.foc[,1])),1,0)
}
within.ref=cbind(person.within.ref,out.within.ref)
within.foc=cbind(person.within.foc,out.within.foc)
between.ref=cbind(person.between.ref,out.between.ref)
between.foc=cbind(person.between.foc,out.between.foc)
outcome.within=rbind(within.ref,within.foc)
outcome.between=rbind(between.ref,between.foc)
outcome=cbind(outcome.within,outcome.between)
}
```