August 2014

DIF Analyses in Multilevel Data: Identification and Effects on Ability Estimates

Yao Wen
University of Wisconsin-Milwaukee

DIF ANALYSES IN MULTILEVEL DATA:

IDENTIFICATION AND EFFECTS ON ABILITY ESTIMATES

by

Yao Wen

A Dissertation Submitted

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in Educational Psychology

The University of Wisconsin Milwaukee

August, 2014

ABSTRACT
DIF ANALYSES IN MULTILEVEL DATA:
IDENTIFICATION AND EFFECTS ON ABILITY ESTIMATES

by

Yao Wen

The University of Wisconsin Milwaukee, 2014
Under the Supervision of Professor Cindy M. Walker

Fairness is an important issue in educational testing in that different groups of examinees should have equal probabilities of answering an item correctly, provided they have the same capabilities. Therefore, differential item functioning (DIF) analyses were developed due to the possibility of bias in cognitive or achievement tests. Data are multilevel structured in educational testing as students are nested within teachers who are nested within schools, and which may further be nested within districts. Although DIF analyses have been discussed for decades, they are rarely investigated in multilevel data. In this study, DIF analyses in multilevel data were investigated via a simulation study with an emphasis on studying DIF at the teacher-level only and at both student and teacher levels, followed by the impacts of DIF on ability estimation. The multilevel Rasch models were used to detect DIF at different locations in both exploratory and confirmatory manners. Type I error rates were all accepted at the 0.05 level. The power was larger when conducting confirmatory analyses. The magnitude of DIF at both levels and the proportion of manifest groups at both levels were two most influential factors on the power of detecting of DIF. However, no influential factors found had impacts on ability estimates. The interpretation of results, possible reasons, limitations, and further studies were discussed.

TABALE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to Dr. Cindy M. Walker, my advisor, for her support and guidance from the beginning to the end of this work. Her patience and contribution are critical in the completion of this dissertation. Dr. Walker graciously provided me the resources such as access to computers and servers to run my simulation. With her quick responses to my questions and her invaluable feedbacks to my drafts, I was able to complete the whole process in a timely manner. I am also very grateful to my committee members, Dr. Razia Azan, Dr. Daniel Bolt, Dr. Ehsan Soofi, and Dr. Bo Zhang for their insightful and constructive comments.

I am grateful to all my friends who have contributed to the success of this work. Kevin Cappaert gave many comments and suggestions on my design and coding, assisting me by proofreading and editing my dissertation. Leanne Freeman supported and helped me spiritually during the most difficult time in this process.

I would also thank to my dear parents, Mianchang Wen and Caiyun Fu, who always believed in the importance of education in a person's life and supported me in every possible way to get a good education throughout my life. I also appreciate my brother for taking care of our dad when he was in the hospital.

Finally, I am thankful to my beloved husband, Li Yang, who sticks with me all the way. Long distance relationship is hard but he never complains. Rather, he provided everything possible to support me for my study and living. His encouragement when times got rough is much appreciated and duly noted.

CHAPTER 1 Introduction

1.1 Statement of the Problem

Fairness is an important issue in educational testing in that different groups of examinees should have equal probabilities of answering an item correctly, provided they have the same capabilities. Therefore, differential item functioning (DIF) analyses were developed due to the possibility of bias in cognitive or achievement tests. When DIF is present, different groups of individuals have different probabilities of getting a correct answer to an item even if they are of the same ability. The presence of DIF can be a serious problem in educational testing because it can threaten the validity of the test (Thissen, Steinberg, & Wainer, 1988, 1993). Strictly speaking, when biased items appear in a test, DIF should be observed. However, if DIF is observed, it is not necessarily due to item bias; judgmental or statistical follow-up analyses must be conducted to determine the presence of item bias (Zumbo, 1999). Therefore, ability estimation bias can lend some additional evidence when making decisions on whether an item or a test is biased.

Additionally, educational testing data is naturally multilevel because students are nested within classes which are nested within schools which are further nested within districts and states. As a result, multilevel models have received more attention in recent years due to the development of computing power and the availability of new software to fit these complicated models. The main drawback of using single level models when fitting multilevel data is that it leads to inflated Type I error rates and biased parameter estimates (Raudenbush & Bryk, 2001).Under the item response theory (IRT) framework, the unidimensional item response model can simultaneously be viewed as a two-level model such that items are nested within individuals. The person trait, or ability, is

characterized as a random parameter which is intended to facilitate marginal maximum likelihood estimation (MMLE) of item parameters (Harwell, Baker, & Zwarts, 1988). Thus, the person trait can be decomposed at higher levels, incorporating covariates that may affect the person trait. The item trait, on the other hand, is usually treated as fixed effect in the IRT model. When a manifest group covariate is added, the IRT model can be used to detect DIF (Luppescu, 2002). When the data are multilevel, DIF could occur at a higher level, such as the teacher level. For example, teacher effectiveness has been studied for decades because it is believed to impact student performance or achievement and thus would affect the estimate of the person trait (ETS, 2004; Medley, 1977). Therefore, it is reasonable to hypothesize that students with the same abilities would have different probabilities of correctly answering an item due to differences in teaching effectiveness. This hypothesis can be tested by conducting DIF analyses using the multilevel IRT model to locate the source of DIF.

DIF has been studied for decades. In most DIF analyses research, one underlying assumption is that the existence of DIF causes test bias; however, this assumption contains two major flaws. Firstly, DIF is necessary, but not sufficient, condition for differential test functioning (DTF), because of the known impact of cancellation (Shealy & Stout, 1993). The second flaw is that DIF is a necessary, but not sufficient condition, for item bias. This is because, as mentioned previously, if DIF is observed other substantive evidence is needed to determine if DIF is actually item bias. Similarly, the decision that test bias exists should be based on the presence of DTF in conjunction with other statistical or judgmental evidence. For example, a negative impact on ability estimation could provide additional evidence that test bias exists, due to DTF. However,

previous studies have shown that the presence of DIF has little effect on ability estimates or on the use of tests in prediction or selection (Neisser, Boodoo, Bourchard, Boykin, Brody, Ceci, Halpern, loehlin, Perloff, Sternberg, & Urbina, 1996; Roznowski, & Reith, 1999; Sackett, Borneman, & Connelly, 2008; Wells, Subkoviak, & Serlin, 2002)

1.2 The Purpose of the Study

DIF analyses in multilevel data are much more complicated than "just adding one level". Due to the fact that DIF can occur at the student level and/or the teacher level, DIF analyses can be conducted at the student level, the teacher level, or both levels. Previous studies in measurement invariance have indicated that when DIF is present at the teacher level, DIF analyses only need to be conducted at this level since the teacher-level DIF does not vary within clusters (Jak, Oort and Dolan, 2014; Ryu, 2013). When DIF is present at the student level, the situation becomes complicated as student-level intercepts and slopes can be random, and the student-level manifest groups may interact with clusters. The research exploring DIF analyses in multilevel data within an IRT framework is scattered and this study was designed to shed some light upon this issue.

This study focuses on DIF detection and ability estimation in multilevel data in terms of uniform DIF. A simulation study was conducted to investigate whether the proposed multilevel IRT model could locate the source of DIF correctly, whether ability estimates are affected by the presence of DIF and, if so, to what degree. The multilevel Rasch model was adopted to detect DIF. MULTILOG 7.0 was implemented to obtain ability estimates. Sources of DIF were simulated at either the student or teacher levels, or at both teacher and student levels. Based on previous studies (Roznowski & Reith, 1999;

Wellset al., 2002), it is known that the magnitude of DIF and the proportion of DIF items affect ability estimation the most.

1.3 The Significance of the Study

This study explores DIF identification with multilevel data in three different situations. In the first situation, DIF is present only at the student level and it is consistent across teacher level clusters. This situation is what traditional DIF analyses assume to be true. In the second situation, DIF is present at the teacher level and the overall impact of DIF at the student level is negligible. In the third situation, DIF is present at both the student and teacher levels. In this situation student-level manifest groups interact with teacher-level manifest groups. The last two types of DIF scenarios would not be detected by traditional DIF analyses.

This study has practical implications. Although it is important to identify DIF items, it is even more important to determine the impact of DIF on ability estimation. Current research primarily concentrates on DIF detection methods, overlooking the practical impact of the presence of DIF. The presence of DIF itself is not sufficient to draw conclusions about test bias or the validity of a test. Therefore, studying the effect of DIF on ability estimation is crucial, in that it provides additional information about the test and facilitates practitioners' decision making, in terms of the final form of the test.

Few studies have explored ability estimation when DIF items are present in multilevel data. Therefore, this study will also shed light on the impact of DIF for practitioners. If ability estimation is not affected by the presence of DIF then the test can be employed directly. In contrast, if ability estimation is impacted by the presence of DIF then the test will need to be modified.

1.4 Overview of Chapters

Chapter 2 introduces key concepts in this study and describes the related literature. Chapter 3 describes the simulation study, including the research design, the simulated conditions, and the evaluation criterions. Chapter 4 presents the results section in which the simulation results are depicted and discussed. The final chapter summarizes the methods and the results, and discusses limitations and possible future development.

CHAPTER 2 Literature Review

2.1 Differential Item Functioning

Differential item functioning (DIF) refers to an item that displays different statistical properties for different manifest groups after the groups have been matched on a proficiency measure (Angoff, 1993). For example, a problem solving item displays DIF if the probability of male examinees correctly answering the item is higher than the probability of female examinees, after controlling for ability. The manifest groups in DIF analyses are known as the focal group, which has the lower probability of obtaining the correct answer to an item, and the reference group, which has a higher probability of obtaining the correct answer to an item.

DIF analyses emerged due to the belief that cognitive and ability tests were biased against minority examinees. However, item or test bias can be due to multiple facets and DIF analyses only provide statistical evidence that is reliant on item scores and group indicators. Practitioners should be cautious when using the results of DIF analyses to generalize to item or test bias. DIF is evidence of such bias if, and only if, the factor causing DIF is irrelevant to the construct being measured by the test.

One common belief in the literature is that DIF is caused due to the multidimensionality of items (Nandakunmar, 1993; Roussos & Stout, 1996; Shealy & Stout, 1993; Walker, 2011; Zumbo, 1999). Unidimensionality is one of the assumptions for unidimensional item response models which states that only one dimension underlies items in a test. DIF occurs when an item measures more than one dimension and two manifest groups differ on their underlying ability distribution for the non-primary dimension(s) that is measured by the item. In such situations, the non-primary dimension

increases the probability of a correct response to an item for examinees in the manifest group that has a higher underlying ability distribution on the non-primary dimension, even though the item may be primarily measuring the primary dimension. The lack of proficiency for examinees in the manifest group that have a lower underlying ability distribution on the non-primary dimension gives them a disadvantage in terms of solving the item correctly. If manifest groups do not differ in their underlying ability distribution on the non-primary dimension then DIF cannot be observed, even if the item is multidimensional (Ackerman, 1992).

Usually, there are two forms of DIF: uniform DIF and non-uniform DIF (or crossing DIF). Uniform DIF occurs when one group performs better than the other group throughout the ability continuum. This implies that an item is more difficult for one group than another across all levels of ability. Technically, uniform DIF exists when the discrimination is equal across manifest groups, but the difficulty is different across manifest groups. Typically the difficulty of the items is greater for the focal group than the reference group. In contrast, non-uniform DIF occurs when there is a difference between the reference and focal group item characteristic curves discrimination parameter. This type of DIF can also exist when both discrimination and difficulty are different for two groups.

2.2 DIF Detection Procedures

2.2.1 Non-IRT model based Approaches

Traditionally, there are numerous procedures to detect DIF. The Mantel-Haenszel (MH) statistic was applied by Holland and Thayer (1988) in determining DIF. The MH statistic is based on the sum of a series of 2×2 contingency tables in which each table

contains the observed correct/incorrect scores from examines in the reference and focal groups. The MH statistic is the most widely used procedure to detect DIF in practice, because it is easy to understand and compute, provides both a significance test and estimate of the magnitude of DIF, and can be employed when the sample size is small (Millsap, 2011). The major criticism of the MH procedure is the adequacy of using the total score as a substitute for the latent trait (Millsap, 2011).

Another popular DIF detection procedure is to compare a set of nested logistic regression models to test for both uniform and non-uniform DIF (Swaminathan & Rogers, 1990). The full model consists of the person trait (the total score or the ability estimate) and group membership as main effects as well as the interaction between them. The first reduced model omits the interaction term. Through the likelihood ratio test, a significant result indicates that the interaction term provides a significant amount of information above and beyond a model that does not include this term. Therefore non-uniform DIF exists. On the other hand, an insignificant result indicates that the interaction term is not necessary. Therefore, non-uniform DIF is not present. Further, the model can be reduced by excluding the group membership term. By comparing this model and the first reduced model, one can determine whether uniform DIF exists. The main issue with the logistic regression procedure is it does not provide the information about the magnitude of DIF.

DIFPACK is a statistical software package designed for detecting uniform DIF in dichotomous items (SIBTEST), polytomous items (Poly-SIBTEST), and crossing DIF (Cross-SIBTEST; Li & Stout, 1996; Shealey & Stout, 1993). This package is recommended because it is based on the theoretical reason for the occurrence of DIF, which is multidimensionality (Walker, 2011). This method adjusts the means of an item,

in terms of differences in the ability distributions for the reference and focal group or impact, using a two-segment piecewise linear regression correction (Jiang & Stout, 1998). As a result, this approach is more accurate in matching the reference and focal groups than MH and logistic regression methods. The estimates of DIF from SIBTEST can measure the magnitude of DIF of which decisions can be made in terms of small, moderate, and large DIF (Nandakumar, 1993).

2.2.2 IRT Model-Based Approaches

In addition to non-parametric DIF detection approaches, there are quite few parametric DIF detection approaches based on item response theory (IRT). IRT models connect the latent traits, or abilities, to item characteristics, such that the latent trait can be predicted by item traits via a monotonically increasing function called an item response function (IRF) or an item characteristic curve (ICC) (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). IRT provides a theoretically useful way to detect DIF such that DIF can be modeled through the use of estimated item parameters and ability. The assumptions of IRT are helpful in understanding DIF detection procedures. First, the unidimensionality assumption corresponds to the multidimensionality perspective on why DIF occurs. Second, the local independence assumption implies that any pair of items is independent, conditional on ability and is a necessary, but not sufficient, condition for the unidimensionality assumption to be met. Third, the item and sample invariance assumption states the item should not vary across samples, up to a linear transformation, which supports the reason for detecting DIF.

IRT models describe the relationship between item characteristics and person latent traits via a probability function. The probability of obtaining a correct answer to an item can be modeled as

$$P_{ij}\left(Y_{ij}=1\middle|\theta\right)=c_i+\frac{1-c_i}{1+e^{-1.7a_i(\theta_j-b_i)}} \quad (2.1),$$

where θ_j is the person trait, or ability; a_i is the item parameter indicating discrimination; b_i is the item parameter indicating difficulty of the item; and c_i is the item parameter referred to as the pseudo-guessing parameter. The difficulty parameter is defined as the location on the ability continuum where the probability of correct response is $\frac{1+c_i}{2}$. It is also the inflexion point of the ICC (Lord, 1980). The more difficult the item, the further the curve is to the right. The parameter a_i is the slope of the ICC at the inflexion point where $\theta=b$. The pseudo-guessing parameter c_i is the lowest asymptote on the ICC (Hambleton & Swaminathan, 1991).

If an item cannot be answered correctly by guessing, then $c_i=0$. In this case, the 3-PL model is reduced to the 2-PL model:

$$P_{ij}\left(Y_{ij}=1\middle|\theta\right)=\frac{1}{1+e^{-1.7a_i(\theta_j-b_i)}} \quad (2.2).$$

Moreover, if all items can be assumed to have the same discrimination parameter, then the 2-PL model is reduced further to the 1-PL model:

$$P_{ij}\left(Y_{ij}=1\middle|\theta_j,b_i\right)=\frac{1}{1+e^{-1.7a(\theta_j-b_i)}} \quad (2.3).$$

When $a=1$, this 1-PL model is reduced to the Rasch model:

$$P_{ij}\left(Y_{ij}=1\middle|\theta_j,b_i\right)=\frac{1}{1+e^{-(\theta_j-b_i)}} \quad (2.4).$$

Many consider the 1-PL model and the Rasch model to be unrealistic because of the assumption that items are all equally discriminating. These models, however, have very

nice mathematical properties. Therefore, tests modeled using the Rasch model have items of the highest caliber.

Based on the item and sample invariance assumption, one parametric IRT-based DIF detection approach is to compare the differences in item parameter (a_i and b_i) estimates using models fit separately to reference and focal group examinees (Camilli & Shepard, 1994; Lord, 1980). However this approach does not take into consideration true differences in ability, or impact, which may exist between the reference and focal group. A better approach is to conduct likelihood ratio tests to compare a set of IRT models in which the reduced model constrains the item parameter to be invariant across groups (Thissen et al., 1988). This method can be implemented using several software packages such as MULTILOG, BILOG-MG, LISCOMP, SPSS LOGLINEAR, LOGIMO, and BIMAIN (Thissen et al., 1993). In this method, DIF free items are required to match people of equal levels of ability, to control for impact. If the item parameters, for a particular item being tested, are not invariant across groups, then an item is flagged as a DIF item and the next item is tested. Another parametric IRT-based method evaluates how different the area measures of ICCs are, between the reference and focal groups (Raju, 1988; Rudner & Gagne, 2001). An important concern in using this method is how to determine the significance of the difference. Although signed area (SA) and unsigned area (UA) can be calculated to evaluate the effect size of DIF (Penfield & Camilli, 2007), they are not efficient to examine the hypothesis of no DIF. This method also fails to take into account the distribution of ability, thus producing misleading interpretations of the size of the observed DIF for specific groups.

In structure equation modeling framework, the multiple indicator multiple cause (MIMIC) model and the multiple group confirmatory factor analysis procedure (CFA) are two common approaches to detect DIF (Hancock & Mueller, 2013). When using the MIMIC model, the latent trait is predicted by a group membership variable, in addition to the measurement model. The significance of the path between the individual indicator and the group membership variable implies the presence of DIF of that indicator (item). Studies have shown the accuracy of using MIMIC model to detect uniform DIF (Finch, 2005; Wang & Shih, 2010; Woods, 2009). Adding a latent variable interaction, the MIMIC model can also be used to test for non-uniform DIF (Woods & Grimm, 2011). The main issue of using the MIMIC model for DIF detection is that the Type I error rates are high (e. g., Finch 2005; Woods & Grimm, 2011). Alternatively, multiple group CFA has been proposed to test for measurement invariance (Meredith, 1993). Four hierarchical levels of invariance are investigated via four nested models in an order of configural, weak, strong, and strict invariance. Weak invariance corresponds to non-uniform DIF and strong invariance corresponds to uniform DIF. Studies have shown that multiple group CFA performs similar to other DIF detection procedures, in terms of power and Type I error rates. However, some DIF detection procedures perform better when items are dichotomous and multiple group CFA tends to perform better when items are polytomous (Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Raju, Laffitte, & Byrne, 2002; Stark, Chernyshenko, & Drasgow, 2006).

2.2.3 Two Level Multilevel Models for DIF Detection

All multilevel models, even though they may have different formulations, rely on the basic hierarchical modeling technique which assumes at least one

random effect that varies across higher levels of the model. One such formulation, models a standard unidimensional IRT function as a multilevel model, with items nested within persons. Kamata (2001) proposed a hierarchical generalized linear model (HGLM) that is algebraically equivalent to the two-level Rasch model. Following the GLM framework, a logit link function and a linear predictor model (level-1 structural model) is formulated in the two-level formulation of the Rasch model.

The level-1 structural model is the item-level model. For an individual j, the response on the item i can be formulized as

$$\eta_{ij} = \log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_{0j} + \beta_{1j}X_{1ij} + \cdots + \beta_{(k-1)j}X_{(k-1)ij},$$

$$\eta_{ij} = \beta_{0j} + \sum_{q=1}^{k-1}\beta_{qj}X_{qij} \quad (2.5),$$

where β_{0j} is the intercept of the model and β_{qj} is the slope of the model. β_{0j} can be viewed as the expected item effect of item i for person j. X_{qij} is the qth variable for person j. It takes on a value of -1 if $q = i$, and 0 otherwise. β_{qj} can be understood as the deviation from β_{0j}. For item i, since $q = i$, $X_{qij} = -1$. Equation 2.5 can be reduced as

$$\eta_{ij} = \beta_{0j} - \beta_{qj} \quad (2.6),$$

where β_{qj} is the effect of qth variable on log of the odds of getting item i correctly for person j. It can be interpreted as the effect of item i when $q = i$.

Level 2 is the person level and β_{0j} is allowed to vary randomly across persons. However, item effects are not allowed to vary across persons. The person level model is

$$\beta_{0j} = \gamma_{00} + u_{0j},$$

$$\beta_{qj} = \gamma_{q0} \quad (2.7)$$

where γ_{00} and γ_{q0} are the fixed effects for β_{0j} and β_{qj} separately; u_{0j} is the random effect associated with person j and is assumed to be a normal distribution of $N(0, \tau_{00})$. u_{0j} can be viewed as the ability of person j.

Combining Equation 2.6 and 2.7, we get

$$\eta_{ij} = \gamma_{00} + u_{0j} - \gamma_{q0} \quad (2.8).$$

This can be rewritten so that the probability of getting item i correctly for person j is

$$P_{ij} = \frac{1}{1+e^{-\left[u_{0j}-(\gamma_{q0}-\gamma_{00})\right]}} \quad (2.9).$$

Equation 2.9 is equivalent to the Rasch model (Kamata, 2001). Comparing this equation to Equation 2.5, $u_{0j} = \theta_j$and $\gamma_{q0} - \gamma_{00} = b_i$. u_{0j} is viewed as the ability of person j and $\gamma_{q0} - \gamma_{00}$ is viewed as the item difficulty parameter for item i.

For DIF detection, Luppescu (2002) extended Kamata's two-level Rasch model and conducted a simulation study to see if the extended model could be used to detect DIF. The sample size, magnitude of DIF, and the proportion of examinees in the focal group were considered as design factors in the study. The interpretation of parameters in the two-level model was revised in order to detect and interpret DIF.

Level-1 model in the extended model was the same as Kamata's level 1 model and consisted of a logit link function and a linear predictor model. In the level 2 of the model, the intercept term was allowed to vary randomly across persons, but no attempt was made to predict this variation. Item effects were not allowed to vary across persons. Rather, a group membership dummy variable was added to the model for each item that was to be tested for DIF. With one DIF item, the level-2 model was formulated as

$$\beta_{0j} = \gamma_{00} + u_{0j},$$

$$\begin{cases} \beta_{1j} = \gamma_{10} + \gamma_{11} G_{1j} \\ \beta_{2j} = \gamma_{20} \\ \quad\quad\vdots \\ \beta_{qj} = \gamma_{q0} \end{cases} \quad (2.10),$$

where G_{qj} is the dichotomous group membership, coded as 1 for the focal group and 0 for the reference group. γ_{q0} can be interpreted as the item difficulty for each item. γ_{q1} is the coefficient associated with each of the dummy variables and can be interpreted as the magnitude of DIF for each item. γ_{00} is the average ability across all examinees and u_{0j} is the deviance of ability from an individual examinee to the average ability.

Luppescu (2002) calculated the root mean squared error (RMSE) to compare the precision of using the Rasch model for DIF detection (Luppescu, 1993) and the extended multilevel Rasch model for DIF detection. Both models performed similarly. The RMSE for the extended multilevel Rasch model was small when the sample size was large, when the magnitude of DIF was small, and when the proportion of people in the focal group was small. However, the Rasch model provided better estimates when the sample size was large. Beretvas and Walker (2011) distinguished DBF from a testlet effect using the multilevel IRT model. They decomposed the DIF into an item-level component and a testlet-specific component. Their simulation study showed that the multilevel IRT model out performed SIBTEST in terms of the identification of DIF, impact, and differential testlet functioning.

Since Kamata's model is restricted to the Rasch model, Swanson, Clauser, Case, Nungester, and Featherman (2002) generalized the logistic model procedure to a hierarchical logistic regression model so that uniform DIF and non-uniform DIF could be detected simultaneously. The level-1 (item level) model in this generalized model was the same as the first reduced model in the logistical regression model procedure, except the

intercept and the slopes were modeled as random across level-2 (person level) clusters. At the person level, the coefficient associated with the level-1 group membership can be predicted by characteristics that may explain DIF. Swanson et al. (2002) demonstrated that the hierarchical logistic regression model could be used as an alternate parameterization method for the 2PL IRT model: The level-1 intercept equals $-a_i b_i$ and the first level-1 slope equals a_i when ability is normally distributed with a mean of zero and standard deviation of one. Although Swanson et al. concluded, via simulation studies, that the hierarchical logistic regression model can be used to successfully investigate the possible causes of DIF; the particular DIF items and the magnitude of DIF are difficult to be determined.

Using the logistic mixed model is yet another way in which one can evaluate test items for DIF (Van den Noortgate & De Boeck, 2005). In contrast to Kamata's multilevel Rasch model, items are treated as random samples from a certain population which implies that the logistic mixed model is based on a model with random item effects (Van den Noortgate, De Boeck, & Meulders, 2003). Using corresponding group membership as covariates, this procedure can identify DIF at the person level or at even higher levels (Van den Noortgate & De Boeck, 2005). If the variance of the random item effects is larger than zero, then DIF exists for at least one item. In this case, empirical Bayes estimates of random item effects for each item can be obtained, to determine which item is functioning differentially. Although the logistic mixed model is flexible, since group membership can also be a random effect, it is well-known that the empirical Bayes estimates are biased. Therefore, detecting DIF for specific items using this framework is particularly challenging.

2.2.4 Three Level Multilevel Models for DIF Detection

In educational testing, nested data, with students nested within classrooms, are frequently encountered. If the researcher is interested in the relationship between student and teacher variables, then the use of traditional models, such as regression models, is problematic and can lead to biased parameter estimates (i.e., Kamata, 2001; Raudenbush & Bryk, 2001). The assumption of independent observations is violated due to the nested data structure. Therefore, multilevel models have been developed to take into account the hierarchical structure. In these models, the variance components are decomposed into each sampling level so that the homogeneity of students in the same class or school can be modeled. Most multilevel models discussed in the last section can be generalized to three-level models, incorporating teacher or school level characteristics that may cause DIF.

Kamata (2001) generalized the two-level Rasch model to the three-level Rasch model. Level 1 is the item level, as it is in the two-level model (Equation 2.5). It is written as

$$\eta_{ijt} = \log\left(\frac{P_{ijt}}{1-P_{ijt}}\right) = \beta_{0jt} + \beta_{1jt}X_{1ijt} + \cdots + \beta_{(k-1)jt}X_{(k-1)ijt},$$

$$\eta_{ijt} = \beta_{0jt} + \sum_{q=1}^{k-1}\beta_{qjt}X_{qijt} \qquad (2.11),$$

where i and j are identical to the level-1 model in the two-level model in Equation 2.5, except for the subscript t that is added to indicate classrooms or teachers. X_{qijt} is the dummy variable that indicates the ith item for person j in classroom t. β_{0jt} is the effect of the reference item and β_{qjt} is the difference between the qth item and the reference item.

Similar to the two-level model, β_{qjt} is constant at the person level. So the person level model for person j in class t is

$$\beta_{0jt} = \gamma_{00t} + u_{0jt},$$

$$\beta_{qjt} = \gamma_{q0t} \quad (2.12),$$

where $u_{0jt} \sim N(v_{00t}, \tau_\gamma)$. This model is identical to the person level model in Equation 2.7, except for the extra subscript t. Here, u_{0jt} indicates the variation of person j within classroom t. The variance of u_{0jt} within class is τ_γ is assumed to be identical for all classrooms. Additionally, γ_{00t} is the effect of the reference item in classroom t; and γ_{q0t} is the effect of the ith item in classroom t.

The overall item effect γ_{00t} can be further modeled at the additional classroom-level. For classroom t, we have

$$\gamma_{00t} = \pi_{000} + e_{00t},$$

$$\gamma_{q0t} = \pi_{q00} \quad (2.13)$$

where $e_{oot} \sim N(0, \tau_\pi)$. At the classroom level, π_{000} and π_{q00} are both fixed item effects; e_{00t} is a random effect with variance τ_π. As in the two-level model, letting $q = i$, a combined model is

$$P_{ijt} = \frac{1}{1+e^{-\left[\left(e_{00t}+u_{0jt}\right)-\left(\pi_{q00}-\pi_{000}\right)\right]}} \quad (2.14).$$

where $\pi_{q00} - \pi_{000}$ is the item difficulty for item i when $q = i$, and π_{000} is the item difficulty for item k. On the other hand, $e_{00t} + u_{0jt}$ can be considered as the ability parameter of person j in classroom t. Unlike the ability term in the two-level model, the ability term in the three-level model contains two random effects. First, e_{00t} is a classroom-level random effect that indicates the average ability of students in classroom t.

Second, u_{0jt} is a person-level random effect of person j in classroom t, implying the size of variation of person j from the average ability of students in classroom t. In a three-level model, ability is decomposed into a person-level ability term and a classroom-level ability term.

Kamata (2001) discussed the impact of person characteristic variables on the estimation of ability using three level Rasch models. From a data demonstration, Kamata concluded that the three-level Rasch model is flexible and can be used to identify a group-characteristic variable that explains variation across higher-level clusters. Furthermore, Kamata, Chaimongkol, Genc, and Bilir (2005) generalized the three-level Rasch model by allowing the coefficient corresponding to the person-level DIF to be random across higher level clusters (schools in their study). That is, the item-level model (Equation 2.11) remains the same, the student-level model becomes

$$\beta_{0jt} = \gamma_{00t} + u_{0jt},$$

$$\begin{cases} \beta_{qjt} = \gamma_{q0t} & \text{if no DIF} \\ \beta_{qjt} = \gamma_{q0t} + \gamma_{q1t} G_{qjt} & \text{otherwise'} \end{cases} \tag{2.15}$$

where G_{qjt} is the group membership at the student level and γ_{q1t} is the effect of DIF. Then the level-3 model becomes

$$\gamma_{00t} = \pi_{000} + e_{00t},$$

$$\begin{cases} \gamma_{q0t} = \pi_{q00} \\ \gamma_{q1t} = \pi_{q10} + e_{q1t} \end{cases} \tag{2.16}$$

where e_{q1t} is the random effect of DIF across schools. If the variance of e_{q1t} is larger than 0, the DIF effect varies across schools. In other words, the effect of the student-level group membership is different from school to school. Jak, Oort, and Dolan (2013)

defined this effect as cluster bias. If cluster bias exists, it is not fair to compare latent means of two groups (Jak et al., 2013).

In Kamata et al.'s (2005) study, they also added group membership to predict the intercept term β_{0jt}, to investigate the impact of DIF. Additionally, they proposed an exploratory DIF model, expanding the level-3 coefficient γ_{q1t} by adding a school-level covariate in order to interpret DIF at the school level. They demonstrated the use of these models and compared the results with the MH procedure, using NAEP data. Although eight items were flagged as DIF items when using the MH procedure, only six of them were detected as DIF items using Equation 2.15 and Equation 2.16 and only two items were detected as DIF using the exploratory DIF model. The potential reason of such discrepancy in results between the MH procedure and Kamata's multilevel Rasch models may be because the models they explored are too complicated.

French and Finch (2010) expanded the hierarchical logistic regression framework to account for DIF at the teacher level. Through a simulation study, they investigated the intraclass correlation coefficient (ICC), the number of clusters, the size of each cluster, DIF magnitude, and DIF location (either the student level or the teacher level). The results were not promising, as power increased as Type I error rates also increased and power decreased as Type I error rates also decreased, even when the model was correct model for the simulated conditions.

Finch and French (2011) discussed the necessity of using multilevel models for nested data via a simulation study. They examined different ICC levels and group memberships at either the student level or the school level. Though multilevel MIMIC models were found to have inflated Type I errors and reduced powers in some conditions,

they recommended using multilevel MIMIC models for better model fit and flexibility in incorporating violators at different levels. Kim, Yoon, Wen, Luo and Kwok (accepted) had similar findings, in which multilevel MIMIC models showed high false positive rates (Type I error rates) even though they could detect DIF when DIF was present at the student level.

In a more recent study, Kim et al. (accepted) introduced multilevel mixture factor models with known classes, to detect uniform and non-uniform DIF with a student-level group membership in multilevel data. Using this model they conducted a series of simulation studies and an empirical data demonstration. The multilevel mixture factor model was used to detect scale-level non-invariance (DTF), while the MIMIC model was used to detect the item-level non-invariance (DIF). They found that both models could be used to successfully identify DIF at the student level; however, multilevel MIMIC models showed relatively moderate to high false positive rates.

The logistic mixed model (Van den Noortgate & De Boeck, 2005) can be easily generalized to a three-level model by incorporating random group effects. These group effects indicated higher level clusters (i.e., schools or classrooms). With a school-level covariate, the logistic mixed model can detect cluster bias as well as school-level DIF. As stated in the last section, the advantage of using this model is that it is flexible because all effects are random; while the predominant criticism of using this model is that it is hard to investigate specific DIF items, in terms of which item shows DIF and to what degree.

2.3 Ability Estimates in DIF Analyses

DIF has been studied for decades with a focus on the identification of DIF and the accuracy of each method. Since there is a general assumption that DIF decreases validity

and causes test bias, common practice dictates that DIF items be rewritten or eliminated from a test. However, by definition, items flagged for DIF cannot indicate either item bias or test bias. Ability estimates can provide substantive information about the impact of DIF. Knowing the influence of DIF on ability estimation, in terms of the number of DIF items and the magnitude of DIF, can help practitioners to make better decisions during the test development process. In spite of the importance of this topic, studies are unexpectedly rare.

Drasgow (1987) investigated measurement bias, in terms of gender and race, using American College Testing (ACT) Assessment Mathematic Usage and English Usage tests. Although item level bias was discovered, some items were biased against the focal group; while other items were biased against the reference group, leading to little evidence of measurement bias for the overall test.

Roznowski (1987) analyzed differences between high school boys and girls in composite test scores which measured a range of topics, some of which were hypothesized to favor boys and some of which were hypothesized to favor girls. Results showed that the correlation coefficients between general intelligence and composite scores were consistent, regardless of group membership. This study provided evidence that items exhibiting group differences do not necessarily indicate poor measurement. Later on, Roznowski and Reith (1999) investigated gender and race differences, using High School and Beyond (HSB) data, with additional regression models using composite scores to predict numerous criteria (i.e., ACT and SAT). Composite scores were created after indexing DIF as biased or not, and indicated group differences. The rank-order correlations were high between composite scores, implying that the order of test scores

were similar, no matter which composite score was used. The correlations between different composite scores and criteria were also similar, indicating test scores were not biased despite group differences. In order to investigate the impact on decision making when using different composite scores, regression models were employed to predict criteria using different composite scores; and t-tests were used to evaluate the slopes from different regression models. Results again showed similar slopes for moderately biased composite scores. Correlations were found to decrease as bias increased and slopes were different for strong biased composite scores in either focal bias or referent bias composites. This study indicated that, with the presence of DIF, the measurement quality is not necessarily degraded; however, the magnitude of DIF may be an influential factor that needs to be considered.

Takala and Kaftandjieva (2000) conducted DIF analyses using a calibration t-test method based on the 1-PL IRT model. They determined test fairness using ability estimates based on four subtest scores: the whole test (40 items), a test with items easier for females (18 items), a test with items easier for males (22 items), and a test with items showing no DIF (29 items). Results indicated that the whole test was not gender biased, regardless if items favored males or females. However, with subtests that included items that only favored males or females, the subtest scores were higher for the favored group.

Similarly, Zumbo (2003) found item-level DIF did not indicate test-level non-invariance, by conducting a simulation study in which the percentage of DIF items (2.9%-42.1%) and the magnitude of DIF (moderate to large) under the CFA framework were considered. The two factors investigated in this study showed no effects on scale

scores. However, Zumbo argued that the presence of item-level DIF may reduce test score in practice, due to an underlying systematic bias.

On the contrary, Pae and Park (2006) investigated the effect of DIF on DTF using CFA by composing 5 subtests of items based on results from IRT-LR procedure: the whole test, items of no DIF, items of balanced DIF, and items of male DIF and female DIF. From the analyses of Korean College Scholastic Aptitude Test (KCSAT) data, they found that item level DIF may influence test level performance, because no cancellation was found with the balanced DIF subtest. They stated that the relationship between DIF and DTF is much more complex than they had expected. The hypothesized reason that no cancellation was found because only uniform DIF was detected, which is not representative of many empirical data sets.

Wells, Subkowviak, and Serlin (2002) investigated the effect of item parameter drift on ability estimation. Item parameter drift occurs when item parameters are not invariant over different testing occasions (Goldstein, 1983). DIF analyses can examine changes in item parameters across occasions. From this study, simulated conditions of item parameter drift had a small effect on ability estimates. Ability estimates were most influenced by the percentage of drifted items, the magnitude of drift, and the test length.

The same findings were obtained in Walker, Zhang, Banks, and Cappaert's (2012) study. A simulation study was conducted which manipulated the number of items containing DIF in a bundle, the test length, and the magnitude of uniform DIF. Results indicated the ability estimates had an inverse relationship with the magnitude of DIF, and the proportion of DIF items.

CHAPTER 3 Methods

In previous studies, student-level DIF and teacher-level (or school-level) DIF have typically been investigated separately (e.g., Finch & French, 2011; Kim et al., accepted; Ryu, 2014). No previous research has considered the presence of DIF at both levels, or the influence of DIF at higher levels on the detection of DIF at lower levels. Therefore, with the assumption of no cluster bias, this study focuses on 1) the detection of DIF when DIF occurs at the teacher level, 2) the detection of DIF when DIF occurs at both levels, where DIF that occurs at higher levels has an impact on DIF that occurs at lower levels, and 3) the effect of the presence of DIF at both levels on ability estimation. Specifically, the following research questions will be addressed:

1. When DIF occurs at the higher level (e.g., teacher level), multilevel Rasch models with the teacher-level covariates would correctly detect teacher-level DIF;

2. When DIF occurs at both teacher and student levels, multilevel Rasch models with both student- and teacher-level covariates would correctly detect DIF at both levels;

3. The magnitude of student-level DIF and the proportion of student-level manifest group would affect the detection of student-level DIF;

4. The magnitude of teacher-level DIF and the proportion of teacher-level manifest group would affect the detection of teacher-level DIF;

5. Factors in 3 and 4 would affect the detection of DIF at both student and teacher levels;

6. The magnitude of DIF and the proportion of DIF items would have an effect on the ability estimation.

Three multilevel Rasch models were explored in this study with the expectation that when DIF occurs at the teacher level, the multilevel Rasch model with a teacher-level covariate should perform best; when DIF occurs at both levels and teacher-level DIF has an effect on student-level DIF, two multilevel Rasch models – one with covariates at both levels and the other one with covariates at both levels and an interaction term – were investigated. Both models should correctly detect DIF, though the model with interaction should perform better in terms of reflecting the effect of teacher-level DIF on student-level DIF. Ability estimation may be influenced by the proportion of DIF items and the magnitude of DIF items.

3.1 Research Design

The design factors were selected based on the purpose of the study. Previous studies have determined that the number of clusters can influence power and Type I error rates in DIF detection (Finch & French, 2011; Kim et al., accepted). Specifically, a greater number of clusters results in larger power and smaller Type I error rates. Therefore, in this study, only one cluster size was selected, 100, that was large enough to have adequate power and Type I error rates. The proportion of DIF items and the magnitude of DIF have been found to have an effect on ability estimation (e.g., Walker et. al., 2013; Wells et al., 2002; Zumbo, 2003). Therefore, these factors were considered in the current study. To manipulate the proportion of DIF items, the number of DIF items varied while the test length remained constant.

The cluster size of the simulated classrooms was based on the reality that classrooms typically contain approximately 30 students. From previous studies, un-equivalent sample sizes in focal and reference group threatens the power in DIF analyses

(Broer, Lee, Rizavi & Powers, 2005; Mazor, Clauser & Hambleton, 1992; Paek & Guo, 2011; Zieky, 1993). Thus, in this study, at student and teacher levels both a balanced and unbalanced design were considered in terms of the manifest groups.

The simulation thus resulted in the following design factors: 2 DIF locations (teacher level and both student and teacher levels) × 2 magnitudes of student-level DIF (0.5, and 0.8) × 2 magnitudes of teacher-level DIF (0.5, and 0.8) × 4 number of teacher-level DIF items (5, 10, 15, and 20 items) × 2 student-level proportions of manifest groups (0.5/0.5 vs. 0.2/0.8) × 3 teacher-level proportions of manifest groups (0.44, 1.00 and 1.44) standard deviation below and above the mean). Other factors that were controlled in this study included the test length (40 items), the number of student-level DIF items (5), the number of clusters (100), and the cluster size (30). A total of 100 replications of each of the 192 conditions was simulated resulting in 19200 data sets.

3.2 Data Generation

Teacher-level ability (the average of student-level ability) was generated from a standard normal distribution, and student-level ability was generated from a multivariate normal distribution with mean of teacher-level ability and variance of an identity matrix. This was done to ensure that there was variability at the student level that could be explained by the teacher level, without which there would be no need for a multilevel model. These actual values of ability were used, in conjunction with simulated item parameters, to generate data. Data were generated using a Rasch model to compute the probability of a particular examinee obtaining the correct answer to an item and comparing it to a random number that was generated from a uniform distribution. If the probability obtained from the Rasch model was greater than, or equal to, the random

number drawn from the uniform distribution the simulated examinee received a score of one (for a correct item). On the other hand, if the probability obtained from the Rasch model was less than the random number drawn from the uniform distribution the simulated examinee received a score of zero (for an incorrect item).

Item difficulty parameters were generated from a standard normal distribution. Student-level DIF items were generated by adding the magnitude of DIF (0.5 or 0.8) to the difficulty parameters of the focal group examinees (e.g., females) while the reference group examinees (e.g., males) difficult parameter remained unchanged. At the teacher level, DIF was generated based on the idea that a teacher's effectiveness might impact performance on particular items, causing them to function differentially, and also might mitigate the impact of any student level DIF that existed. Specifically, teacher effectiveness ratings were generated from a standard normal distribution and used as cut-off values to categorize teachers into three groups: highly effective teachers, average teachers, and non-effective teachers. Three different values of cut-off levels were considered: 0.44 (balanced), 1.00 (unbalanced), and 1.44 (extremely unbalanced). These values were chosen because they controlled the proportion of teachers in each category. As Figure 3.1 illustrates, in the balanced design, using a cut-off value of 0.44 resulted in 33% of teachers being categorized as effective, 34% of teachers being categorized as average, and 33% being categorized as ineffective. On the other hand, in the extremely unbalanced design, using a cut-off value of 1.44 resulted in 7.5% of teachers being categorized as effective, 85% of teachers being categorized as average, and 7.5% of teachers categorized as ineffective.

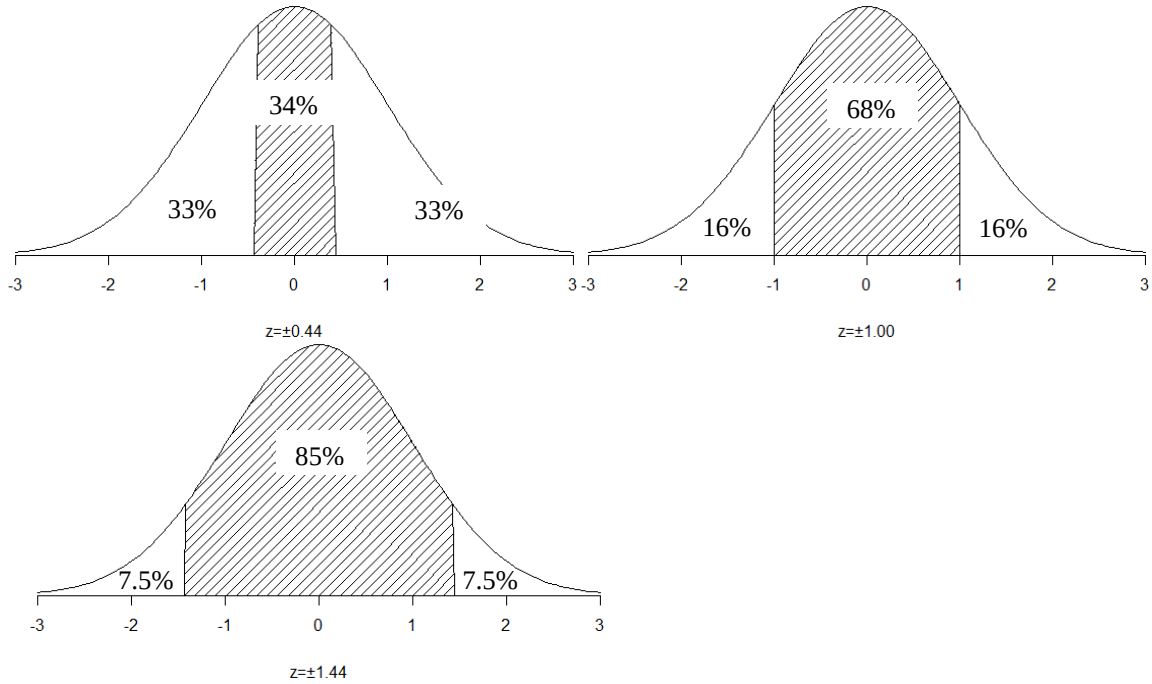

Figure 3.1. Three levels of teacher effectiveness based on different cut-off values

When DIF only occurred at the teacher level, the difficulty parameter for reference group examinees that were taught by average teachers was generated from $b \sim N(0,1)$. For DIF items, the difficulty parameter for students with non-effective teachers was modified as $b + DIF_t$ while the difficulty parameter for students with highly effective teachers was modified as $b - DIF_t$.

Table 3.1

Generating multilevel DIF items

	Manifest groups					
	Non-effective teachers		Average teachers		Effective teachers	
	Reference group	Focal group	Reference group	Focal group	Reference group	Focal group
Teacher level only	$b + DIF_t$		b		$b - DIF_t$	
Teacher-student levels	$b + DIF_t$	$b + DIF_s + DIF_t$	b	$b + DIF_s$	$b - DIF_t$	$b + DIF_s - DIF_t$

When DIF was present at both the student and teacher level, items that functioned differentially at the student level were made more difficult for students that were taught

by ineffective teachers, less difficult for students that were taught by highly effective teachers, and were not changed for students that are taught by average teachers. Table 1 depicts the way in which teacher level DIF was simulated and how it was used to influence student level DIF when both student and teacher level DIF were present.

3.3 Models in This Study

In this study, we assumed items were nested within students and further nested within teachers. In other words, person effects are random, reflecting a random term in the equation. Item effects, on the other hand, are fixed, showing no random term in the equation.

Kamata's (2001) three-level Rasch model was extended and used in this study because adding covariates, at either the person level or the teacher level, allows one to get a better understanding of factors that affect DIF detection or the impact. The biggest advantage of using Katmata's three-level IRT model is that it can estimate item parameters while identifying DIF for multiple items and estimating impact simultaneously. It has been shown that the three-level IRT model can improve the estimation of the relationship between latent traits and predictor variables (Pastor, 2003).

For the three-level model, the student-level group membership or the teacher-level group membership may influence the item effect. There are three situations in terms of the source of DIF. DIF can occur at the student level only, at the teacher level only, and at both student and teacher levels. This study focuses on situations that DIF occurs at teacher only and both student and teacher levels.

When DIF occurs at the teacher level, for example, teachers may influence students' understanding of problems, or approaches to problem solving, students with an

effective teacher may understand the question better or employ better problem solving strategies, so some test items may be easier for students with effective teachers than for students with less effective teachers. Moreover, the observed average student performances with effective teachers may be higher. As a result, the ability estimates of students with effective teachers may be higher. The characteristics of teachers may impact the identification of DIF and influence ability estimation. Therefore, by adding one grouping variable to the item effect at the teacher level, the detection of DIF at the teacher level can be achieved.

The item-level model is the same as Equation 2.11, however, the fixed effect of the intercept is fixed to be zero in practice. Therefore, all items can be freely estimated as no reference group is needed.

$$\eta_{ijt} = \log\left(\frac{P_{ijt}}{1-P_{ijt}}\right) = \beta_{0jt} + \beta_{1jt}X_{1ijt} + \cdots + \beta_{kjt}X_{kijt},$$

$$\eta_{ijt} = \beta_{0jt} + \sum_{q=1}^{k} \beta_{qjt}X_{qijt} \quad (3.1).$$

Here in Equation 3.1, β_{0jt} indicates ability and β_{qjt} indicates difficulty, which are different from the interpretation of Kamata's model in Equation 2.11. The student-level model then becomes

$$\beta_{0jt} = u_{0jt},$$

$$\begin{cases} \beta_{1jt} = \gamma_{10t} \\ \quad \vdots \\ \beta_{qjt} = \gamma_{q0t} \end{cases} \quad (3.2).$$

The teacher-level model is the same as Equation 2.13 except a categorical variable, indicating group membership, is added to the item effect model. Therefore, with one DIF item, the teacher-level model is

$$\begin{cases} \gamma_{10t} = \pi_{100} + \pi_{101} B_t \\ \gamma_{20t} = \pi_{200} \\ \quad \vdots \\ \gamma_{q0t} = \pi_{q00} \end{cases} \quad (3.3),$$

where B_t is the group membership at the teacher level and π_{101} is the effect of group membership at the teacher level, indicating DIF. The rest of the coefficients are the same as the coefficients in Equation 2.13. The combined model of detecting teacher-level DIF is

$$\eta_{ijt} = u_{0jt} + \sum_{q=1}^{k} \pi_{q00} X_{qijt} + \pi_{101} B_t X_{qijt} \quad (3.4).$$

If no hypothesis is made about DIF items, one can run an exploratory analysis by assuming all items should show DIF. Then the combined model becomes

$$\eta_{ijt} = u_{0jt} + \sum_{q=1}^{k} (\pi_{q00} + \pi_{q01} B_t) X_{qijt} \quad (3.5).$$

When DIF occurs at both the student and the teacher level, with the assumption of no interaction between the student-level group membership and clusters, two possible scenarios may occur. One scenario is that the student-level membership has no relationship with the teacher-level group membership. The model is then a synthesis of the model with student-level covariate only and the model with the teacher-level covariate only. With one DIF item at the student level and the teacher level, the model is

Level 1: $$\eta_{ijt} = \beta_{0jt} + \sum_{q=1}^{k} \beta_{qjt} X_{qijt},$$

Level 2: $$\beta_{0jt} = u_{0jt},$$

$$\begin{cases} \beta_{1jt} = \gamma_{10t} + \gamma_{111} G_j \\ \beta_{2jt} = \gamma_{20t} \\ \quad \vdots \\ \beta_{qjt} = \gamma_{q0t} \end{cases},$$

Level 3:
$$\begin{cases} \gamma_{10t} = \pi_{100} + \pi_{101} B_t \\ \gamma_{20t} = \pi_{200} \\ \quad \vdots \\ \gamma_{q0t} = \pi_{q00} \end{cases} \quad (3.6).$$

And the combined model is

$$\eta_{ijt} = u_{0jt} + \sum_{q=1}^{k} \pi_{q00} X_{qijt} + \gamma_{111} G_j X_{1ijt} + \pi_{101} B_t X_{1ijt} \quad (3.7).$$

This model can be used to test the hypothesis that DIF occurs both at the student and teacher levels independently. If no hypothesis made about DIF items, the model for the exploratory model is

$$\eta_{ijt} = u_{0jt} + \sum_{q=1}^{k} (\pi_{q00} + \gamma_{q11} G_j + \pi_{q01} B_t) X_{qijt} \quad (3.8).$$

The other scenario that may occur is that the student-level membership may interact with the teacher level group membership, as described in the previous section. In this scenario the item level model is still the same as Equation 2.11 and the student level model remains the same; however, the slope of the group membership can be modeled at the teacher level. The resulting with one DIF item then becomes:

Level 1:
$$\eta_{ijt} = \beta_{0jt} + \sum_{q=1}^{k} \beta_{qjt} X_{qijt},$$

Level 2:
$$\beta_{0jt} = u_{0jt},$$

$$\begin{cases} \beta_{1jt} = \gamma_{10t} + \gamma_{11t} G_j \\ \beta_{2jt} = \gamma_{20t} \\ \quad \vdots \\ \beta_{qjt} = \gamma_{q0t} \end{cases},$$

Level 3:
$$\begin{cases} \gamma_{10t} = \pi_{100} + \pi_{101} B_t \\ \gamma_{20t} = \pi_{200} \\ \quad \vdots \\ \gamma_{q0t} = \pi_{q00} \end{cases}$$

$$\gamma_{11t} = \pi_{110} + \pi_{111} B_t \quad (3.9),$$

where γ_{11t} is a random slope that can be modeled at the teacher level and π_{111} is the coefficient displaying the effect of teacher-level characteristic (e.g., teacher effectiveness) on the student-level group membership (e.g., gender). The student-level random slope can be understood as a variable at the teacher level which has different effects on male and female students. Since item parameters are fixed effects, no residual terms are included for β_{qjt}, γ_{q0t}, and γ_{11t}. The combined model is

$$\eta_{ijt} = u_{0jt} + \sum_{q=1}^{k} \pi_{q00} X_{qijt} + \pi_{110} G_j X_{1ijt} + \pi_{101} B_t X_{1ijt} + \pi_{111} G_j B_t X_{1ijt} \qquad (3.10).$$

In Equation 3.10, the term $\pi_{111} G_j B_t X_{1ijt}$ is also called the cross-level interaction because G_j is a student-level covariate and B_t is a teacher-level covariate. In Equation 3.10, the interaction should be interpreted when it is significant. The main effects (π_{110} and π_{101}) should be interpreted when the interaction is not significant. Similarly, the model for the exploratory analyses is

$$\eta_{ijt} = u_{0jt} + \sum_{q=1}^{k} (\pi_{q00} + \pi_{q10} G_j + \pi_{q01} B_t + \pi_{q11} G_j B_t) X_{qijt} \qquad (3.11).$$

In this study, both exploratory and confirmatory DIF analyses were explored and compared to learn the appropriateness of using the proposed multilevel Rasch models in both exploratory and confirmatory DIF analyses.

In summary, three models were used to identify DIF using multilevel data: (1) The three-level Rasch model, with only a teacher-level covariate (ML-teacher), depicted in Equations 3.4 and 3.5; (2) The three-level Rasch model with independent covariates at both levels (ML-Both), depicted in Equations 3.7 and 3.8; and (3) The three-level Rasch model with a cross-level interaction (ML-Inter), depicted in Equations 3.10 and 3.11. These three models were explored in both exploratory and confirmatory DIF analyses. In the confirmatory case, a group membership variable was only included for the five items

(from 5 to 20 items at the teacher level) simulated to function differentially. In the exploratory case, a group membership variable was included for all items. All models were estimated and evaluated using PROC GLIMMIX in SAS 9.4.

3.4 Estimation Method

The likelihood function, used for estimation, can be expressed as

$$L = \prod_{j=1}^{N} \prod_{i=1}^{I} {P_{ij}}^{y_{ij}} (1 - P_{ij})^{1-y_{ij}} \quad (3.12).$$

In which case the log likelihood is

$$l = y_{ij} \log\left(P_{ij}\right) + \left(1 - y_{ij}\right) \log\left(1 - P_{ij}\right) \quad (3.13).$$

By integrating out the random effects (the person parameter), the marginal log likelihood function can be used to obtain item parameter estimates, as well as ability estimates, using an iterative procedure (e.g., EM; Bock & Aitkin, 1981).

The PROC GLIMMIX procedure in SAS provides three different estimation options. The default estimation procedure used derives an approximation to the marginal likelihood and its partial derivatives, using linearization techniques. Breslow and Clayton (1993) use the term quasi-likelihood to describe this method. Given starting values for unknown parameters, the first order Taylor series is used to linearize the logistic function, leading to a standard linear mixed model. Suppose $\boldsymbol{Y}$ represents the $(n \times 1)$ vector of response data and γ is a $(r \times 1)$ vector of random effects. The generalized linear mixed model is

$$E(\boldsymbol{Y}|\gamma) = g^{-1}(\boldsymbol{X}\beta + \boldsymbol{Z}\gamma) = g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu} \quad (3.14)$$

where $g(\cdot)$ is a differentiable monotonic link function (logistic in this study) and $g^{-1}(\cdot)$ is its inverse. The matrix $\boldsymbol{X}$ is a $(n \times p)$ matrix of rank k, and $\boldsymbol{Z}$ is a $(n \times r)$ design matrix for the random effects. The random effects are assumed to be distributed as

$\gamma \sim N(0, G)$ and the variance of response data is assumed to be $\text{var}(\boldsymbol{Y}|\gamma) = \boldsymbol{A}^{1/2}\boldsymbol{R}\boldsymbol{A}^{1/2}$ where $\boldsymbol{A}$ is a diagonal matrix and contains the variance functions of the model and $\boldsymbol{R}$ is a variance matrix which can be specified by the user through the RANDOM statement (SAS Institute Inc., 2013). With a first-order Taylor series of μ about $\tilde{\beta}$ and $\tilde{\gamma}$ (Wolfinger & O'Connell, 1993), the model becomes

$$g^{-1}(\eta) \doteq g^{-1}(\tilde{\eta}) + \tilde{\boldsymbol{\Delta}}\boldsymbol{X}\left(\beta - \tilde{\beta}\right) + \tilde{\boldsymbol{\Delta}}\boldsymbol{Z}(\gamma - \tilde{\gamma}) \tag{3.15}$$

where $\tilde{\boldsymbol{\Delta}} = \left(\frac{\partial g^{-1}(\eta)}{\partial \eta}\right)_{\tilde{\beta},\tilde{\gamma}}$ is a diagonal matrix of derivatives of the conditional mean evaluated at the expansion locus. After combining this with Equation 3.14 and rearranging Equation 3.15, the model can be rewritten as

$$\boldsymbol{P} = \boldsymbol{X}\dot{\beta} + \boldsymbol{Z}\dot{\gamma} + \epsilon \tag{3.16}$$

where $\boldsymbol{P} = \boldsymbol{X}\beta + \boldsymbol{Z}\gamma$ and $\boldsymbol{X}\dot{\beta} + \boldsymbol{Z}\dot{\gamma} + \epsilon = \tilde{\boldsymbol{\Delta}}^{-1}\left(\boldsymbol{Y} - g^{-1}(\tilde{\eta})\right) + \boldsymbol{X}\tilde{\beta} + \boldsymbol{Z}\tilde{\gamma}$. This model becomes a linear mixed model with pseudo-response $\boldsymbol{P}$, fixed effects $\dot{\beta}$, random effects $\dot{\gamma}$, and $\text{var}(\epsilon) = \text{var}(\boldsymbol{P}|\dot{\gamma})$.

After obtaining Equation 3.16, parameters are estimated using estimation methods for linear mixed models, and the estimates are used for a new Taylor series of the logistic function. The fixed effects are estimated through a marginal quasi-likelihood (MQL) procedure and the random effects are estimated through a penalized quasi-likelihood (PQL) procedure.

Although the quasi-likelihood procedure is effective, some researchers have reported that it yields underestimates for both fixed effects and variance components for dichotomous data (e.g., Goldstein & Rasbash, 1996; Rodriguz & Goldman, 1995). Another disadvantage of this method is due to the use of quasi-likelihood procedure.

Model fit statistics based on this likelihood function are approximate and may not be used for evaluating model fit (Hox, 2002).

Laplace's method for integral approximation is an alternative approach that can be used to approximate the likelihood function. This method expands the exponent of the integrand, expressed as a function of the random effects in a second-order Taylor series around the maximizer of the exponent function, and uses normal theory to find the integral (Skrondal & Rabe-Hesketh, 2004). Laplace estimates typically show better asymptotic behavior and less small-sample bias than the quasi-likelihood method. However, Laplace estimation is based on the conditional independence assumption and, thus, requires no random residual (R-side) covariance structure (Wolfinger, Tobias, &Sall, 1994). Adaptive quadrature is yet another method that can be used to integrate over the random effects distribution. If the distribution is assumed to be normal, Gause-Hermite quadrature can be used to approximate the integral by a weighted sum of the integrand, evaluated as the specified number of quadrature points (Hedeker & Gibbons, 2004).

In a small pilot study conducted, with several design factors explored in this study, these three estimation methods yielded comparable estimates of parameters. Therefore, the quasi-likelihood estimation was used in this study, due to its efficiency.

3.5 DIF Detection Procedure

When using multilevel Rasch models, statistically significant coefficients associated with group membership are indicative of DIF. Specifically, in this study there were two manifest groups at the student level and three manifest groups at the teacher level. So, at the teacher level, teacher effectiveness was dummy coded as B_{t1} and B_{t2},

indicating effective teachers and non-effective teachers correspondingly. In the ML-Teacher model, π_{q01} indicates teacher-level DIF (Equation 3.5); in the ML-Both model, γ_{q11} indicates student-level DIF and π_{q01} indicates teacher-level DIF (Equation 3.8); and in the ML-Inter model, π_{q11} indicates the effect of teacher-level DIF on student-level DIF. The significance of parameter estimates was evaluated at $\alpha = .05$ level. Type I error rates and power were calculated to evaluate how well the tested models identified DIF.

3.6 Parameter Recovery

To evaluate how well the multilevel Rasch models were performing, estimates of difficulty were compared to the true values. Bias, correlation and the root mean square error (RMSE) between estimated parameters (difficulty) and true parameters were calculated. Bias is the deviation between the estimated parameters and the true parameters such that:

$$Bias = \frac{\sum_{i=1}^{N}(\hat{b}_i - b_i)}{Nn} \quad (3.17),$$

where $\hat{b}$ is the estimated parameter, b is the true parameter, N is the number of replications in the simulation study, and n is the sample size. Bias is used to evaluate the distance from the estimated value to the true values as well as the direction. The root mean square error (RMSE) is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{b}_i - b_i)^2}{Nn}} \quad (3.18).$$

RMSE is used to evaluate the absolute magnitude of difference between estimated parameters and true parameters. The correlation, on the other hand, is used to evaluate the rank order between the estimated and true parameters.

3.7 Ability Estimation

In order to investigate the influence of DIF on ability estimation, MUTILOG 7.0 was used to obtain ability estimates. One condition was added to this study to enable the comparisons of the results obtained from the newly proposed three-level models to the model that would typically be used in practice. Specifically, Kamata's two-level model (Rasch model) was fit to the three level data, for all of the conditions explored, and the bias, RMSE and correlation were calculated for this condition for comparison purposes.

Analysis of variances (ANOVAs) were conducted, given that a full factorial design was used. The effect size

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} \tag{3.19}$$

of each combination of the design factors was used as a way to determine practical significance. Given a mixed design, let A be a between-subject factor, B be a within-subject factor and S be subjects, the effect size of a between-subject factor is (Maxwell, & Delaney, 2004)

$$\omega_A^2 = \frac{SS_A - df_A(MS_{S/A})}{SS_A + SS_{S/A} + MS_{S/A}} \tag{3.20};$$

the effect size of a within-subject factor is

$$\omega_B^2 = \frac{df_B(MS_B - MS_{B\times S/A})}{SS_B + SS_{B\times S/A} + SS_{S/A} + MS_{S/A}} \tag{3.21};$$

and the effect size of a within-subject interaction is

$$\omega_{AB}^2 = \frac{df_A df_B(MS_{AB} - MS_{B\times S/A})}{SS_{AB} + SS_{B\times S/A} + SS_{S/A} + MS_{S/A}} \tag{3.21}.$$

Effects that explained more than five percent of the total variance were investigated descriptively.

CHAPTER 4 Results

4.1 DIF Detection

The rates of inadmissible solutions were examined across simulation conditions for both the exploratory and confirmatory analyses. An inadmissible solution refers to the non-convergence of a model. For both sets of analyses, inadmissible solution rates were zero or near zero for most cases. Across all conditions, the highest rate of inadmissible solutions was 3.89% for the exploratory analyses, while the highest rate of inadmissible solutions was 1.27% for the confirmatory analyses.

Type I error rates were evaluated at the 0.05 level, for both exploratory and confirmatory analyses (Table 4.1). Type I error refers to the false detection of an invariant item as non-invariant, in this study. That is, the estimates of π_{q01} in Equation 3.5 and 3.8 and π_{q11} in Equation 3.9 are significant for DIF free items. According to Bradley (1978), the acceptable range of Type I error rates is computed with a formulae $\alpha \pm 1/2\alpha$. When $\alpha = 0.05$, the Type I error rates between .025 and .075 are considered reasonable.

Average Type I error rates across all non-DIF items are presented in Table 4.1. As depicted in Table 4.1, the column indicates the location of simulated DIF and the row indicates the models used to obtain the estimates of DIF. For both the exploratory and confirmatory analyses, the Type I error rates for the three-level Rasch models with covariates at the teacher level only and at both levels fell within Bradley's range (Table 4.1). No significant factors were found to explain differences in Type I error when using ANOVA analyses to determine if any of the factors studied influenced the Type I error rate, for both exploratory and confirmatory analyses (See Table A.1 and Table A.2 in

Appendix A). For the confirmatory analyses with DIF items known, it was not possible to make a Type I error so no estimates are produced (NA – power in Table 4.1). In other words, for example, in one condition five teacher-level DIF items were simulated and fit using the ML-Teacher model with covariates on those five items only. Thus, the detection of DIF is power rather than Type I error. This is also true when using ML-Both and ML-Inter models to detect DIF generated at the teacher level levels when fitting a confirmatory model. However, ML-Both and ML-Inter models can also detect student-level DIF, leading to type I errors when DIF is simulated at the teacher level only. Regardless of the location of DIF, the magnitude of DIF, the percentage of DIF items, and the sample size in each manifest group, the three-level Rasch models exhibited acceptable Type I error rates in both exploratory and confirmatory analyses.

Table 4.1

Type I Error Rates

Models	Data generation			
	Teacher level only		Both levels	
	Exploratory	Confirmatory	Exploratory	Confirmatory
ML-Teacher	0.053	NA - power	0.054	NA – power
ML-Both	0.053	0.047	0.052/0.053	NA – power
ML-Inter	0.049	0.045	0.048	NA - power

One purpose of this study was to investigate how well the multilevel Rasch model performed in terms of DIF detection. Therefore, power was evaluated across all conditions to investigate how well each model performed. Power is defined as the proportion of cases in which DIF items were correctly detected. Any value equal or larger than 0.8 was presumed to be indicative of high power.

Average power, across all conditions and items, is presented in Table 4.2. Overall, when exploratory DIF analyses were conducted, the power depended on the location of

DIF, the magnitude of DIF, and the proportion of teachers in each manifest group. For the exploratory DIF analyses, the power associated with using ML-Both models to test for student level DIF was 0.812, when DIF was present at both levels. However, when trying to detect teacher-level DIF, the power of ML-Teacher and ML-Both was only 0.511 (or 0.510) no matter if DIF was present at only the teacher level or at both levels. The power of using ML-Inter model decreased to 0.442, on average, across all conditions. On the contrary, conducting confirmatory DIF analyses yielded almost perfect results (power ≥ 0.80).

Table 4.2

Power

Model	Data generation			
	Teacher level only		Both levels	
	Exploratory	Confirmatory	Exploratory	Confirmatory
ML-Teacher	0.511	0.954	0.504	0.953
ML-Both	0.511	0.948	0.812/0.510	1.000/0.953
ML-Inter	0.442	0.802	0.551	0.913

In order to evaluate the effects of the design factors on power, full factorial ANOVA analyses were employed in terms of DIF location when using ML-Teacher and ML-Both models; and mixed ANOVA analyses were employed when using the ML-Inter model due to the interest of within-subject effects. Factors that were associated with a large effect size (η^2 or $\omega^2 > 0.05$) obtained are presented in the rest of this chapter. The power obtained from conducting exploratory DIF analyses is depicted in Figure 4.1 to Figure 4.4; while the power obtained from conducting confirmatory DIF analyses is presented in Figure 4.5 to Figure 4.8. Results from the ANOVA analyses are presented in Appendix A for the sake of conciseness in the Results section.

For both the exploratory and confirmatory analyses, the results obtained from fitting the ML-Teacher and ML-Both model to detect teacher level DIF are presented together because they were similar. This is followed up by the results obtained when fitting the ML-Inter model in an exploratory manner. And at the last, the results obtained when fitting the ML-Inter model when conducting a confirmatory analysis are presented. There were no significant effects when fitting ML-Teacher and ML-Both models in a confirmatory manner. Therefore, these results are not interpreted. Thus, to begin with, the power of detecting student-level DIF when using the ML-Both model is discussed.

As depicted in Table A.3 (in Appendix A), when using the ML-Both model to detect student-level DIF and conducting exploratory analyses, only the magnitude of student-level DIF had an effect on power ($\eta^2_{S_{DIF}} = .216$). As shown in Figure 4.1, when student-level DIF = 0.5, the power was only 0.668; whereas when student-level DIF = 0.8, the power was 0.972.

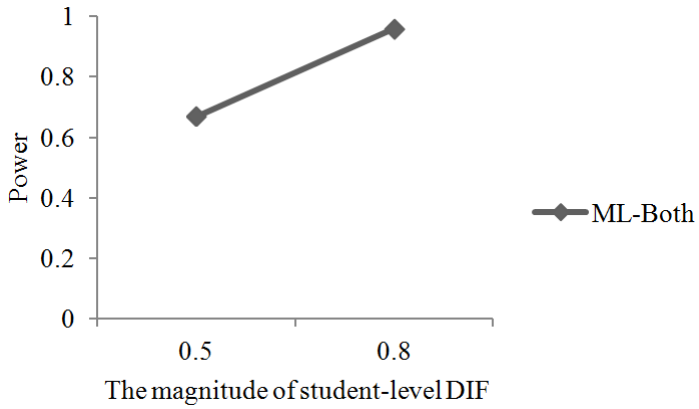

Figure 4.1 Power of Student-level DIF When Conducting Exploratory Analyses

The results obtained from the full factorial ANOVA, when exploratory DIF analyses were conducted to detect teacher-level DIF, indicated that two main effects (the magnitude of teacher-level DIF and the proportion of teachers in each category) had large effects on power ($\eta^2_{T_{DIF}} = .318$ and $\eta^2_{T_{group}} = .201$ in Table A.4 in Appendix A). ML-

Teacher and ML-Both models yielded similar power when detecting teacher-level DIF. Figure 4.2 was chosen to exhibit the effect because it includes more information. As depicted in Figure 4.2, power increased as the magnitude of teacher-level DIF increased; power decreased as the teacher-level grouping design factor went from balanced (SD = 0.44) to extremely unbalanced (SD = 1.44). In the worst condition the power was only 0.221. This occurred when the magnitude of teacher-level DIF was 0.5 and the teacher-level grouping design factor was extremely unbalanced. In the best condition the power reached 0.842. This occurred when the magnitude of teacher-level DIF was 0.8 and the teacher level grouping design factor was balanced.

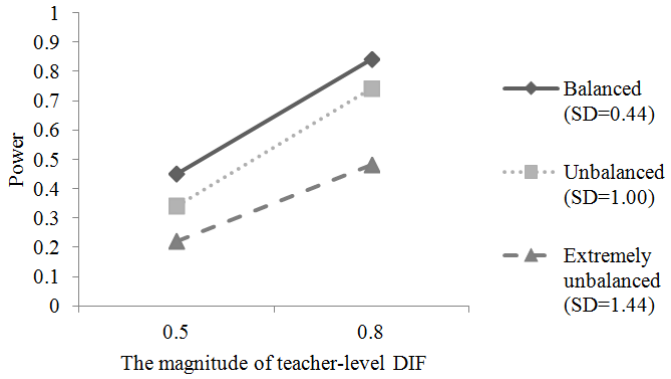

Figure 4.2 Power of Detecting Teacher-level DIF When Conducting Exploratory Analyses

In short, with a large magnitude of DIF and balanced teacher-level manifest groups, the power of DIF detection using either ML-Teacher or ML-Both model is high, even when conducting exploratory analyses. Practitioners may use these models to detect either teacher-level DIF or both student and teacher-level DIF when it is hypothesized that the magnitude of DIF is large and the teacher-level manifest groups are balanced. On the other hand, the ML-Inter model detects student-level and teacher-level DIF integratedly. Therefore, as shown in Table 4.2, the ML-Inter model had lower overall power than the ML-Both model.

Three main factors from the full factorial ANOVA that were most influential on power rates, when fitting the ML-Inter model in an exploratory manner, were the magnitude of DIF at the teacher levels, the proportion of teachers in each manifest group, and the location of DIF ($\eta^2_{T_{DIF}} = .266$, $\eta^2_{T_{group}} = .169$, $\eta^2_{DIF_{Loc}} = .054$; Table A.5). On average, power increased as the magnitude of teacher-level DIF increased (0.375 when teacher-level DIF = 0.5 and 0.527 when teacher-level DIF = 0.8). Power decreased as the teacher-level manifest group design went from the balanced to the extremely unbalanced (0.518 in the balanced design, 0.484 in the unbalanced design, and 0.355 in the extremely unbalanced design). Power increased when simulated DIF was at both levels (0.518), as opposed to only at the teacher level (0.442). Because all power was low for the significant main effects that were obtained, additional analyses were conducted separately in terms of DIF location.

Mixed ANOVA analyses were conducted to find most influential design factors when DIF occurred either at the teacher level only or at both the teacher and student levels. The between factors considered were the following: the magnitude of student-level DIF; the magnitude of teacher-level DIF; the proportion of student-level manifest groups; and the proportion of teacher-level manifest groups. The within factors considered were the following six student-teacher manifest groups: male students with average teachers, male students with effective teachers, male students with non-effective teachers, female students with average teachers, female students with effective teachers and female students with non-effective teachers. Due to dummy coding, male students with average teachers were the reference category in the ML-Inter model. If male students were treated as the reference category, then one would be testing for student-

level DIF; if average teachers were treated as the reference category, then one would be testing for teacher-level DIF. However, in our case, when treating male students with average teachers as the reference category one is testing for DIF at both the teachers and student levels.

When conducting exploratory DIF analyses to detect teacher-level DIF only using the ML-Inter model, the within factors considered were the following: female students with effective teachers; female students with non-effective teachers; male students with effective teachers; and male students with non-effective teachers. In this case students with average teachers are the reference category and thus are not discussed. Results from the mixed ANOVA (Table A.6 in Appendix A) indicated that the within factors did not differ across the four levels. However, the magnitude of teacher-level DIF, the proportion of teacher-level manifest groups and their interaction, all of which are between factors, were found to be influential in impacting the power ($\omega^2_{T_{DIF}} = .351$, $\omega^2_{T_{group}} = .261$, and $\omega^2_{T_{DIF} \times T_{group}} = .062$). A similar pattern to what was described earlier was found. Specifically, the power increased as the magnitude of teacher-level DIF increased; and the power decreased as the teacher-level manifest group design changed from balanced to extremely unbalanced. Specifically, when teacher-level DIF = 0.5, the power was low in all manifest groups (0.383 in balanced design, 0.293 in unbalanced design and 0.198 in extremely unbalanced design). When teacher-level DIF = 0.8, the power was still low in unbalanced and extremely unbalanced design (0.385 and 0.641). However, the power reached 0.778 in the balanced design.

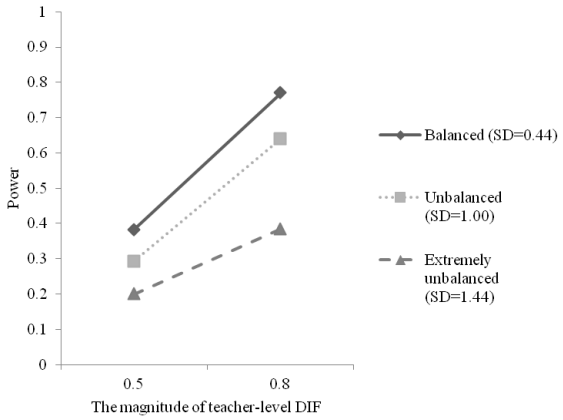

Figure 4.3 Power of Detecting Teacher-level DIF using ML-Inter When Conducting Exploratory Analyses

When DIF was present at both the student and teacher levels and exploratory DIF analyses were conducted using the ML-Inter model, the results of the mixed ANOVA analysis indicated that there were three interactions that were found to have a large effect size, in terms of the within-subject factors. As shown in Table A.7, which can be found in the appendix, the following were found to have a large effect size: the interaction between the student-teacher manifest group and the magnitude of student-level DIF ($\omega^2_{S_{DIF}\times ST_{group}} = .054$); the interaction between the student-teacher manifest group and the magnitude of teacher-level DIF ($\omega^2_{T_{DIF}\times ST_{group}} = .053$); and the interaction between the student-teacher manifest group and the proportion of teacher-level manifest groups ($\omega^2_{T_{group}\times ST_{group}} = .068$). Three main effects were also found for the between-subject factors: the magnitude of student-level DIF ($\omega^2_{S_{DIF}} = .097$), the magnitude of teacher-level DIF ($\omega^2_{T_{DIF}} = .165$), and the proportion of teacher-level manifest groups ($\omega^2_{T_{group}} = .086$). The three interaction effects are presented in Figure 4.4.

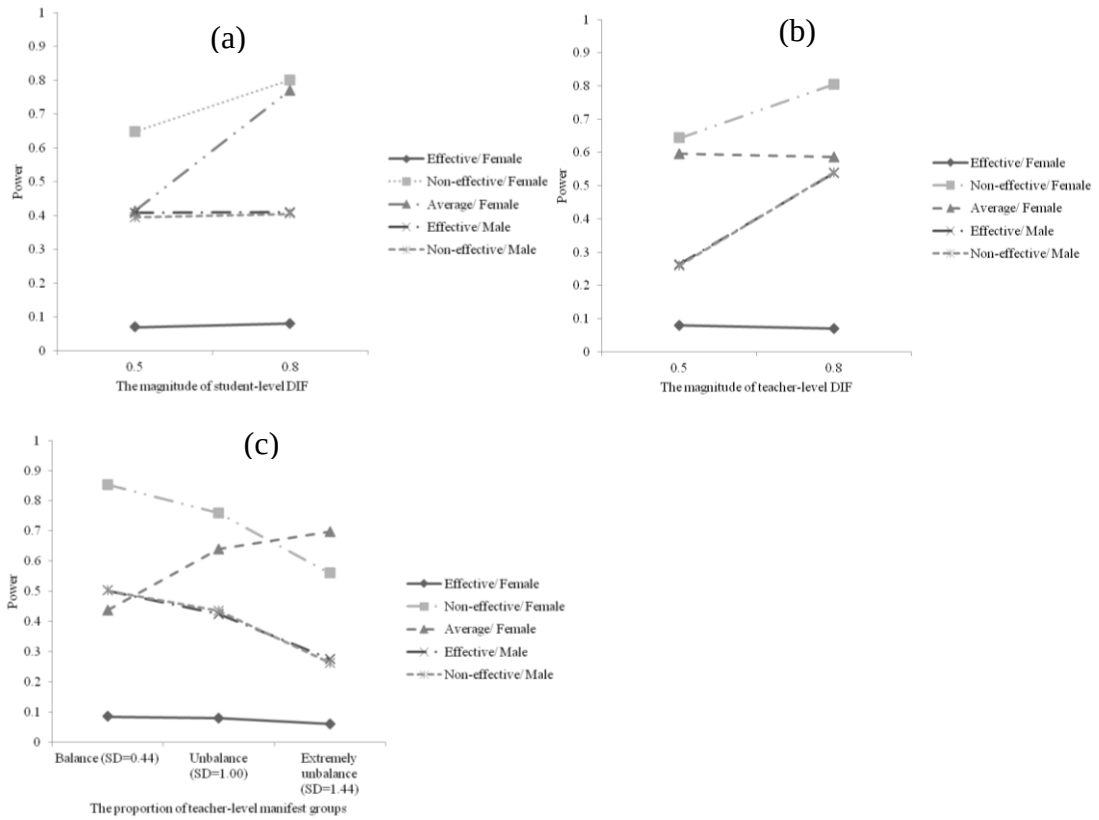

Figure 4.4 Power of Detecting Both-level DIF using ML-Inter When Conducting Exploratory Analyses

Recall that when DIF occurred at both levels, the final DIF effect was a combination of student-level DIF and teacher-level DIF. Therefore, when student-level DIF and teacher-level are equivalent, final DIF is canceled out for females (focal group examinees) with effective teachers (e.g., $-DIF_t + DIF_s = -0.5 + 0.5 = 0$). When student-level DIF and teacher-level are not equal, final DIF is not cancelled out, but remains very small (e.g., $-DIF_t + DIF_s = -0.5 + 0.8 = 0.3$). As a result, the detection of such small DIF at both levels was not very good with power ranging from only 0.06 to 0.14 (the bottom line in Figure 4.4). In Figure 4.4 (a), one can see that the power of detecting DIF at the both levels for each student-teacher manifest group increased as the magnitude of student-level DIF increased for female students with non-effective teachers and average teachers; while the power remained the same for male students with effective

and non-effective teachers. This may be because the magnitude of student-level DIF only has an effect on detecting student-level DIF. The power for females with non-effective teachers was relatively large (from 0.648 to 0.824) because the simulated DIF is the sum of student- and teacher-level DIF ($DIF_t + DIF_s$), which resulted in a large magnitude of DIF which ranged from 1.00 to 1.60. For females with average teachers, the power increased from 0.413 to 0.789. This may be due to having a large proportion of teachers in the average group in the extremely unbalanced design.

In Figure 4.4 (b), one can see that the power of detecting DIF at both levels, increased as the magnitude of teacher-level DIF increased for females with effective and non-effective teachers, as well as for males with effective and non-effective teachers. This may be because the magnitude of teacher-level DIF only impacts the detection of teacher-level DIF. Once again, power was relatively large for females with non-effective teachers (from 0.643 to 0.815). However the power of detecting DIF at both levels for males with effective or non-effective teachers was not large even when the magnitude of DIF = 0.8 (power = 0.538).

In Figure 4.4 (c), one can see that the power of detecting DIF at both levels decreased as the teacher-level manifest group design changed from a balanced design to an extremely unbalanced design, except for females with average teachers. This is because the proportion of teachers in the average group became larger when the teacher-level manifest group moved from a balanced design to an extremely unbalanced design. The largest power was obtained for females with non-effective teachers in a balanced design (0.853).

To summarize, conducting exploratory DIF analyses with the proposed models and multilevel data largely depends on the magnitude of DIF, the location of DIF and the proportion of teachers in each manifest group. Overall, the power was not promising when conducting exploratory analyses. On the contrary, conducting confirmatory DIF analyses, without any model misspecification, yielded almost perfect results (Table 4.2). Results from the ANOVA found that there were influential design factors when DIF was generated at both student and teacher levels using the ML-Inter model. These results are presented next.

As depicted in Table A.8, when a confirmatory approach was taken and the ML-Inter model was used four within-subject interactions were found to impact power rates: a three-way interaction was found between the student-teacher manifest group, the magnitude of student-level DIF and the magnitude of teacher-level DIF ($\omega^2_{S_{DIF}\times T_{DIF}\times ST_{group}} = .111$); a two-way interaction between the student-teacher manifest group and the proportion of teacher-level manifest group ($\omega^2_{T_{group}\times ST_{group}} = .131$); a two-way interaction between the student-teacher manifest group and the proportion of student-level manifest group ($\omega^2_{S_{group}\times ST_{group}} = .080$); and a two-way interaction between the student-teacher manifest group and the magnitude of teacher level DIF ($\omega^2_{T_{DIF}\times ST_{group}} = .103$). In addition, two between-subject interactions and two between-subject main effects were found to impact power rates: the interaction between the magnitude of student-level DIF and the teacher-level DIF ($\omega^2_{S_{DIF}\times T_{DIF}} = .120$); the interaction between the magnitude of teacher-DIF and the proportion of teacher-level manifest groups ($\omega^2_{T_{group}\times T_{DIF}} = .109$); the main effect of the magnitude of teacher-level

DIF ($\omega^2_{T_{DIF}} = .258$); and the main effect of the proportion of teacher-level manifest groups ($\omega^2_{T_{group}} = .256$). Interactions are interpreted below.

The three-way within-subject interaction is shown in Figure 4.5. For female students with effective teachers, no DIF was simulated when the magnitude of student-level DIF was equal to the magnitude of teacher-level DIF. Therefore, the power of detecting both-level DIF (or Type I error) was around 0.05. When the magnitude of student-level DIF did not equal the magnitude of teacher-level DIF, the power of detecting both-level DIF was about 0.323. For other manifest groups, the power of detecting both-level DIF increased as the magnitude of DIF increased. When teacher-level DIF = 0.5, the power of detecting both-level DIF was lower for male students with effective or non-effective teachers than for female students with effective or non-effective teachers (0.793 vs. 0.977). However, when the magnitude of teacher-level DIF = 0.8, the power of detecting both-level DIF was high for all manifest groups.

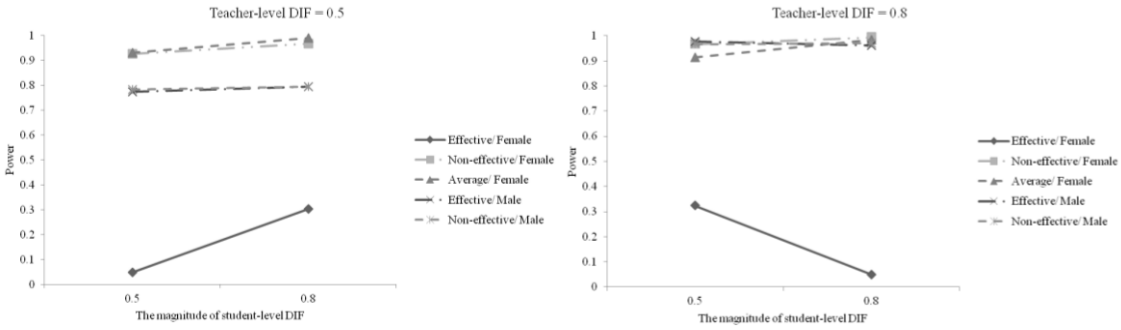

Figure 4.5 Power of Detecting Both-level DIF When Conducting Confirmatory Analyses: 3-way interaction

The interaction between the student-teacher manifest group and the proportion of student-level manifest groups is presented in Figure 4.6. In the balanced design where 50% of the students were in the focal group (female) and 50% of students were in the reference group (male), the power of detecting both-level DIF was as high as 0.834. In

the unbalanced design where only 20% of the students were in the focal group and 80% of the students were in the reference group, the power of detecting both-level DIF was low for female students with effective and non-effective teachers; but high for male students with effective and non-effective teachers. The results indicate that the proportion of student-level manifest groups also has an effect on the power to detect DIF at the teacher-level, when average teachers are used as the reference category. This is why power is so low (0.483) for female students with average teachers.

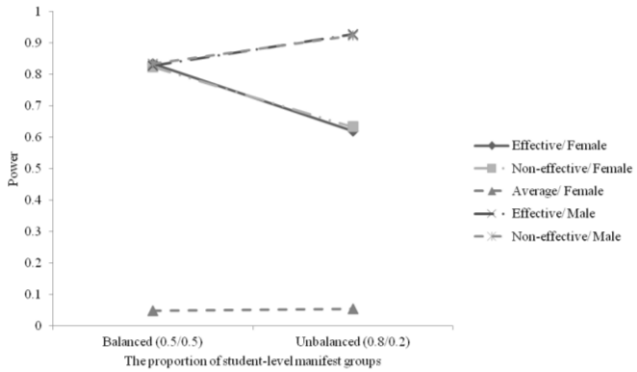

Figure 4.6 Power of Detecting Both-level DIF When Conducting Confirmatory Analyses: S_group

The interaction between the student-teacher manifest groups and the proportion of teacher-level manifest group is presented in Figure 4.7. As stated previously, the power of detecting both-level DIF obtained was low for female students with effective teachers due to the small magnitude of DIF (0 or 0.3) that was simulated. For other conditions, when the proportion of teachers in each group changed from a balanced design to an extremely unbalanced design, the power decreased, except for female students with average teachers. However, even with this decrease in power, the power of detecting both-level DIF was as high as 0.758 in the worst condition, which was for male students with effective or non-effective teachers. Female students with non-effective teachers yielded largest power (from 0.901 to 1.000) due to the large magnitude of simulated DIF.

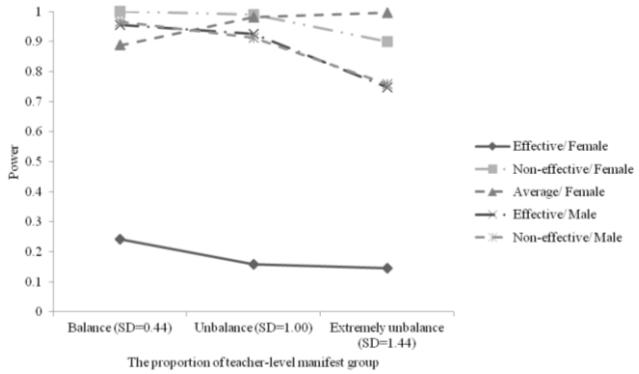

Figure 4.7 Power of Detecting Both-level DIF When Conducting Confirmatory Analyses: T_group

The interaction between the magnitude of teacher-level DIF and the proportion of teacher-level manifest groups is presented in Figure 4.8. When the magnitude of teacher-level DIF was small (0.5), the power of detecting both-level DIF was relatively small only in the extremely unbalanced design (0.754). When the magnitude of teacher-level DIF was large (0.8), the power of detecting both-level DIF was large in all conditions (from 0.946 to 0.985).

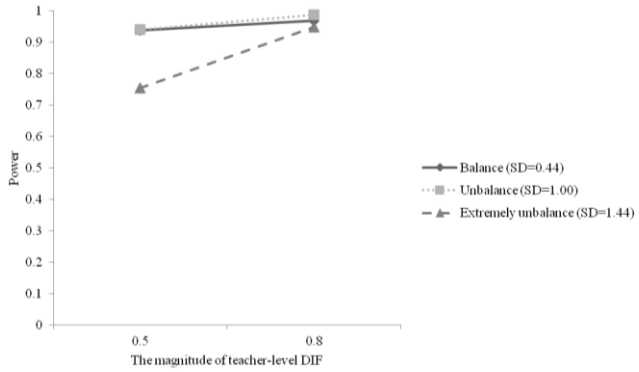

Figure 4.8 Power of Detecting Both-level DIF When Conducting Confirmatory Analyses: T_group × T_DIF

In summary, confirmatory analyses were found to be better than exploratory analyses in terms of smaller Type I error rates and larger power, while the magnitude of DIF and the proportion of either students or teachers in each manifest group had the greatest influence on the detection of DIF when conducting confirmatory analyses.

4.2 Parameter Recovery

The difficulty parameter is a fixed effect in the multilevel Rasch model. In both exploratory and confirmatory DIF analyses, the difficulty parameter estimates were very close to the true parameter. The correlation coefficients were nearly one, bias was small and RMSE's were also small (Table 4.3). No factors were found to have effect sizes larger than 0.05.

Table 4.3

Bias, correlation and RMSE of difficulty parameter

	Exploratory			Confirmatory		
	Correlation	bias	RMSE	Correlation	bias	RMSE
ML-Teacher	1.00	-0.02	0.20	1.00	0.01	0.11
ML-Both	1.00	-0.00	0.20	1.00	-0.00	0.10
ML-Inter	1.00	-0.00	0.20	1.00	-0.00	0.11

4.3 Ability Estimates

Ability estimates were obtained using MULTILOG 7.0 using a 1-PL model to fit the multilevel data. A baseline condition was added such that data were generated from a Rasch model and ability was estimated using the same Rasch model so that the results obtained from the multilevel data could be compared to a best case scenario. Comparing results from the simulated data to this baseline condition allows for a better understanding of the factors that have an impact on ability estimation.

Table 4.4 depicts the correlation, bias and RMSE for all conditions, including the Rasch No DIF condition. As the table illustrates, regardless of the magnitude of DIF, the number of DIF items at the teacher level, or the level at which DIF occurred, the bias was always near zero and the correlation was always high (0.96). The only difference observed in the table is that the RMSE's were noticeably smaller for the Rasch NO DIF

condition (0.59 vs. 0.35) than for the other conditions. However, no significant factors were found to influence ability estimates. Therefore, when using the Rasch model to estimate ability when DIF is present in multilevel data, the standard errors of ability estimates will be biased, but not to a great extent. These findings were not entirely consistent with previous studies.

Table 4.4

Bias, correlation and RMSE for ability estimates

DIF Location	Teacher-level DIF	Student-level DIF	Number of DIF items	Correlation	bias	RMSE
Rasch No DIF	0	0	5_item	0.94	0.00	0.35
			10_item	0.94	0.00	0.35
			15_item	0.94	0.00	0.35
			20_item	0.94	0.00	0.35
		0	5_item	0.94	0.00	0.35
			10_item	0.94	0.00	0.35
			15_item	0.94	0.00	0.35
			20_item	0.94	0.00	0.35
	0	0	5_item	0.94	0.00	0.35
			10_item	0.94	0.00	0.35
			15_item	0.94	0.00	0.35
			20_item	0.94	0.00	0.35
		0	5_item	0.94	0.00	0.35
			10_item	0.94	0.00	0.35
			15_item	0.94	0.00	0.35
			20_item	0.94	0.00	0.35
Teacher level	0.5	0	5_item	0.96	0.00	0.59
			10_item	0.96	0.01	0.59
			15_item	0.96	0.00	0.59
			20_item	0.95	0.00	0.60
		0	5_item	0.96	-0.01	0.58
			10_item	0.96	0.00	0.59
			15_item	0.96	0.00	0.59
			20_item	0.95	0.00	0.60
	0.8	0	5_item	0.96	0.00	0.59
			10_item	0.96	-0.01	0.59
			15_item	0.95	0.00	0.60
			20_item	0.95	0.00	0.61
		0	5_item	0.96	0.00	0.58

DIF Location	Teacher-level DIF	Student-level DIF	Number of DIF items	Correlation	bias	RMSE
			10_item	0.96	0.00	0.59
			15_item	0.95	0.00	0.60
			20_item	0.95	0.00	0.61
Both level	0.5	0.5	5_item	0.96	0.00	0.59
			10_item	0.96	0.00	0.59
			15_item	0.96	-0.01	0.59
			20_item	0.95	0.00	0.60
		0.8	5_item	0.96	-0.01	0.59
			10_item	0.96	0.00	0.59
			15_item	0.96	0.00	0.59
			20_item	0.95	0.01	0.60
	0.8	0.5	5_item	0.96	0.00	0.59
			10_item	0.96	0.00	0.59
			15_item	0.95	0.00	0.60
			20_item	0.95	0.00	0.61
		0.8	5_item	0.96	0.00	0.59
			10_item	0.96	0.00	0.59
			15_item	0.95	0.00	0.60
			20_item	0.95	-0.01	0.61

CHAPTER 5 Discussion

This study investigated the use of multilevel Rasch models for the detection of DIF with multilevel data under a variety of research conditions. Overall, DIF in multilevel data is a complicated issue, due to the existence of different types of random effects. This study explored DIF in multilevel data with the invariant item assumption in IRT, as well as fixing the student-level group membership across clusters. This assumption helps to simplify the detection of DIF.

5.1 DIF Detection

In this study, using a multilevel Rasch models proved to be successful in identifying DIF in multilevel data, when using a confirmatory approach, at both the student and teacher level. In traditional DIF analyses, it is typically assumed that DIF is due to characteristics that are only manifest at the student level. This presumes, in a multilevel modeling framework, that the impact of DIF is the same across all clusters. By definition, DIF can also occur at the teacher level. Using the example in this study, effective teachers employ better instructional methods, or tools, to help students with problem solving. After a period of time, students with effective teachers may show better performance even though the students may have the same ability level as students in classrooms with less effective teachers. In a multilevel situation such as this, a researcher may be interested in investigating DIF at the teacher level in order to understand the differential performance among students. The existence of teacher-level DIF should not influence DIF detection at the student level, if teacher-level characteristics do not vary within clusters (Ryu, 2013). This has been verified through this simulation study.

In this study, the ML-Teacher model was showed to detect teacher-level DIF successfully. As expected, the magnitude of teacher-level DIF and the proportion of teacher-level manifest group had effects on the detection of teacher-level DIF when conducting an exploratory analysis. With large magnitude of teacher-level DIF and equal proportion of teacher-level manifest group, the ML-Teacher model showed high power even in the exploratory analyses (Figure 4.2).

The ML-Both model can be used to detect both-level DIF separately as student-level DIF and teacher-level DIF. When teacher-level DIF only occurs, using the ML-Both level model yielded comparable power with the ML-Teacher model and acceptable Type I error rates (Table 4.2). As expected, the magnitude of student-level DIF, the magnitude of teacher-level DIF, and the proportion of teacher-level manifest group had effects on the detection of both-level DIF in the exploratory analyses (Figure 4.1 and Figure 4.2). However, the proportion of student-level manifest group did not impact the detection of student-level DIF.

The ML-Inter model, on the other hand, can be used to detect DIF integratedly when DIF occurs at both student and teacher levels, indicating DIF for each student-teacher manifest group. Moreover, as stated in Chapter 4, if the student-level reference group (e.g., male students) is treated as the reference category, the results indicate the student-level DIF; if the teacher-level reference group (e.g., average teachers) is treated as the reference category, the results indicate the teacher-level DIF; and if the student-teacher reference group (e.g., male students with average teachers), the results indicate the integrated both-level DIF. Again, as expected, in this study, the magnitude of student-level DIF, the magnitude of teacher-level DIF, the proportion of student-level manifest

group, and the proportion of teacher-level manifest group did have effects on the detection of both-level DIF.

If a researcher is interested in testing for DIF at both levels, but is interested in the effects of DIF at each of those levels, the detection of DIF can be achieved in two ways: (1) Using one model with group membership covariates at each level (the ML-Both model); or, (2) Using two models, one with student-level group membership as a covariate and a second one with teacher-level group membership as a covariate (the ML-Teacher model). In this study, similar results were found for the three-level Rasch model with teacher-level covariates (the ML-Teacher model) and the three-level Rasch model with independent covariates at both levels (the ML-Both model) in detecting teacher-level DIF.

When DIF occurs at both student and teacher levels, teacher-level DIF may influence student-level DIF, as an interaction may exist between teacher-level and student-level DIF. For example, if effective teachers introduced a method related to spatial memory to solve a math problem, boys may benefit more than girls. As long as the test does not test spatial memory, but tests how to solve a math problem, the differences in responses between boys and girls are due to DIF. In this situation, the three-level Rasch model with a cross-level interaction (the ML-Inter model) could be used. The differentiation of the ML-Both model and the ML-Inter model was a major focus of the current study. If one asks the question “Is there student or teacher level DIF?”, the ML-Both model is sufficient to answer that question. If one asks the question “Does teacher effectiveness influence student performance in terms of their gender or race?”, the ML-Inter model is more appropriate.

Consistent with previous studies (e.g., Finch, 2005; Walker et al, 2012; Zumbo, 1999), the magnitude of DIF and the proportion of the manifest group was found to affect DIF detection most. More specifically, as Linacre (2013) illustrated, when DIF = 0.5, the smallest sample size for each manifest group must be 300 in order to detect DIF with appropriate power and Type I error rate control. When DIF = 1.0, the sample size requirement greatly decreases, to only 100 persons in each manifest group. In this study, extremely unbalanced design at the teacher level resulted in only 6 to 9 teachers in efficient or non-effective teacher groups, and as few as 180 to 270 students in such groups, the power of detecting teacher-level DIF was far too low. However, the student-level proportion of manifest group was not found to have a profound effect on DIF detection in this study. The reason for this may have been the large sample size of at the student level. Even when only 20% of the students were focal group examinees, this was equivalent to 600 students, which is large enough for DIF detection at the student-level.

5.2 Ability Estimates

Previous studies have indicated that ability estimates are influenced by the percentage of items and the magnitude of DIF (e.g., Walker et al, 2012; Zumbo, 2003). In this study, however, no factors were found to have a significant effect on ability estimation. Regardless of the percentage of DIF items and the magnitude of DIF, the standard errors of ability estimates were large. Moreover, if rank ordering examinees is of interest, the presence of DIF in a hierarchical data structure will not affect this rank ordering of ability estimates at all. In this study, only five, out of forty, items were set up as DIF items at the student level, which is a percentage of only 12.5% of items. Walker et al. (2012) found that having 15% of items that function differentially may lead to

statistically significant ability differences between reference- and focal-group examines. In this study, the percentage of DIF items at the teacher level was not found to influence ability estimation as expected. Since no factors were found to be influential on ability estimation in this study, other evaluation methods may be carried out. One method to test ability estimation is to employ *t*-tests to compare ability estimates between reference- and focal-group students. The other method is that person fit statistics can be used to investigate the misfit in each response pattern. It is hasty to conclude ability estimation will not be affected by DIF in multilevel data based only on the current study. More studies needs to be done, in terms of the impact of the hierarchical structure, the cluster bias, and DIF at each level on ability estimation.

5.3 Practical Implications

In practice, empirical researchers conducting DIF analyses using multi-level models are often concerned about the appropriate model to use, the order in which one should detect student- and teacher-level DIF, and the correct interpretation of the results. As described in Chapter 2, there are numerous procedures to detect DIF in multilevel data. However, most of the previous studies did not consider the order in which one should detect student and teacher level DIF. The multilevel Rasch model is flexible, easy and efficient to apply in SAS. One can add fixed or random effects to test different assumptions. If no hypotheses are made, cluster bias should be detected first, using a random effect for the item of interest across clusters. If the random effect is significant, this implies that the item difficulty varies across clusters. In other words, the item functions differentially from one class to another, if classroom is the third level of the model. The significant random effect violates the invariance assumption of IRT models.

In this case the corresponding analyses should focus on why this happened. This situation is different from what Van den Noortgate and De Boeck (2005) proposed, which was using logistic mixed models and assuming that items are randomly sampled from a population (e.g., item bank). With an insignificant random effect of items, student-level DIF can first be tested, followed by the detection of teacher-level DIF. One can investigate DIF in a stepwise fashion, adding one parameter at a time while checking the significance of estimates as well as model fit if not using quasi-likelihood.

One criticism of the multilevel Rasch model is that it requires a relatively large sample size (Hox, 2002; Raudenbush& Anthony, 2001). Alternatively, it is possible to estimate the multilevel Rasch model using a Markov Chain Monte Carlo (MCMC) simulation. One of the biggest advantages of MCMC is that it works well with small sample sizes (Christensen, Johnson, Branscum, & Hanson, 2010). The weakness of this method is that it is too time consuming.

The multilevel mixture model with known classes (Muthén, 2002) may be another alternative DIF detection procedure. Comparing the multilevel mixture model to the multilevel MIMIC model, the multilevel mixture model detects DIF with high power and acceptable Type I error rates (Kim et al, accepted). The Mixture Rasch model was introduced by Rost (1990, 1991) to identify two latent classes that reflected knowledge states on physical achievement. After that, studies have been employed using the mixture Rasch model to detect "latent DIF," or differential performance due to differential levels of the latent trait. Conceptually, this model is more suitable to detect impact, but it can also be used to detect DIF by using the observed group membership. More studies are

needed to investigate the behavior of the multilevel mixture model under the IRT framework.

5.4 Limitations

The main limitation of this study was the assumption of fixed item effects. Although this assumption is consistent with the item invariance assumption in item response theory, it is not necessarily true in real testing scenarios. As mentioned previously, cluster bias should be tested prior to conducting any DIF analyses. If cluster bias exists, one can still conduct DIF analyses using a multilevel Rasch model with a fixed group membership and random item effects. After controlling for the effect of clusters, one can interpret the results obtained from fitting this model as whether a given characteristic leads to DIF. However, the issue with such a model is that the size of the random effects is hard to determine. Usually, the presence of a random effect is determined when the variance of the random effect is larger than zero. However, with the presence of both random item effects and DIF, the decision about which more affects test performance is unclear.

Another limitation of this study is that only generalized Rasch models were discussed. The Rasch model is famous for its mathematical simplicity; but criticized for its lack of flexibility (restricting the discrimination parameter to one). However, due to the fact that multilevel models are so complicated, generalizing the 2-PL model to multilevel data will be computationally challenging. The current popular methods which account for both discrimination and difficulty include the multilevel MIMIC model and the multilevel mixture factor model with known classes (Kim et al., accepted). In fact, when using these models the discrimination and difficulty parameters in the 2-PL model

can be obtained by transforming the factor loadings and residual variances from these models (Lord & Novick, 1968). Moreover, it is well known that the use of the MIMIC model for DIF detection yields high Type I error rates (Finch, 2005; Kim& Yoon, 2011). In addition, the multilevel mixture factor modeling method with known classes has been shown to perform well in a recent study designed to determine if this procedure could be used to detect student-level DIF in multilevel data (Kim et al., accepted). It is important to note that the multilevel mixture factor model allows student-level factor loadings to vary across clusters. With empirical data, researchers would need to test the random student-level factor loadings first before determining the most appropriate model to use.

5.5 Conclusion

DIF analyses have been conducted for decades, but DIF analyses in multilevel data have not been considered until recently, with the development of the ability to estimate these models which require complex computational techniques. The multilevel Rasch model discussed in this study performed well in detecting DIF at the student or/and teacher level with certain hypotheses about which item would show DIF. The estimates of fixed parameters were close to the true values even with the quasi-likelihood estimation, indicating the multilevel Rasch model is reliable in terms of DIF detection. If more random effects are added into the model, the Laplace estimation or the adaptive quadrature estimation may be used, though they are both time consuming and have restrictions with particular statements in SAS (SAS Institute Inc., 2013). Ability estimates were found to suffer overall, in terms of large standard deviation; but no factors were found to have a significant impact on ability estimation. For future research one might

investigate more evaluation methods or consider more ways to evaluate the impact on ability estimation.

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.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.) *Differential Item Functioning* (pp. 3 - 23). Hillside: Lawrence Erlbaum Associates.

Beretvas, S. N., & Walker, C. M. (2011). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement*, 72(2), 200-223.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, 46, 443-449.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of American Statistical Association*, 88(421), 9-25.

Broer, M., Lee, Y., Rizavi, S., & Powers, D. (2005). *Ensuring the fairness of GRE writing prompts: Assessing differential difficulty* (Research report RR-05-11). Princeton, NJ: Educational Testing Service.

Camilli, G., & Shepard, L. A. (1996). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, 72(1), 19-29.

Educational Testing Service. (2004). *Where we stand on teacher quality: an issue paper from ETS*. Teacher Quality Series.

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

Finch, W. H., & French, B. F. (2011). Estimation of MIMIC model parameters with multilevel data. *Structural Equation Modeling*, 18, 229-252.

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, 47(3), 299-317.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, 20, 369-377.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE publication.

Hancock, G. R., & Mueller, R. O. (2013). Structure Equation Modeling: A Second Course (2nd ed.). *A volume in quantitative methods in education and the behavioral sciences: issues, research, and teaching*. Information Age Publishing: Charlotte, NC.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: a didactic. *Journal of Educational Statistics*, 13, 243-271.

Hedeker, D. R., & Gibbons, R. D. (2004). *Longitudinal Data Analysis*. Unpublished manuscript.

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

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

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: detecting violations of measurement invariance across clusters in multilevel data. *Structure Equation Modeling*, 20(2), 265-282.

Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structure Equation Modeling*, 21(1), 31-39.

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

Kamata, A., Chaimongkol, S., Genc, E., & Bilir, K. (April 2005). *Random-effect differential item functioning across group units by the hierarchical generalized linear model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: a comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, 18, 212-228.

Kim, E. S., Yoon, M., Wen, Y., Luo, W., & Kwok, O. (accepted). Within-level group factorial invariance in multilevel data: Multilevel factor mixture and multilevel MIMIC models, *Structure Equation Modeling*.

Li, H. & Stout, W. (1996). A new procedure for detecting crossing DIF. Psychometrika, 61, 647-677.

Linacre, J. M. (2013). Differential item functioning DIF sample size nomogram, *Rasch Measurement Transactions*, 26(4), 1391.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Nahwah, NJ: Lawrence Erlbaum Associates.

Luppescu, S. (1993). DIF detection examined: which item has the real differential item functioning? *Rasch Measurement Transactions*, 7(2), 285-286.

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

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research*, 45, 975-999.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. Educational and Psychological Measurement, 52, 443-451.

Medley, D. M. (1977). *Teacher competence and teacher effectiveness: a review of process-product research*. Washington D. C.: American Association of Colleges for Teacher Education.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543.

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge.

Muthén, B. O. (1994). Multilevel covariate structure analysis. *Sociological Methods & Research*, 22, 376-398.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement*, 30(4), 293-311.

Neisser, U., Boodoo, G. Bourchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, 51, 77-101.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equalivalence/invariance. *Organizational Research Methods*, 7(4), 361-388.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*(1; ISSU 51), 81-118.

Pae, T., & Park, G. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing*, 23(4), 475-496.

Paek, I., & Guo, H. (2011). Accuracy of DIF estimates and power in unbalanced designs using the Mantel-Haenszel DIF detection procedure. *Applied Psychological Measurement*, 35(7), 518-535.

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: an illustration. *Applied Measurement in Education*, 16(3), 223-243.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Hand Book of Statistics*, Volume 26: *Psychometrics* (pp. 125-167). New York: Elsevier

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

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement evidence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87(3), 517-529.

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: application and data analysis methods* (2ed). Newbury Park, CA: SAGE Publications, Inc.

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, 73-90.

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

Rost, J. (1991). A logistic mixture distribution model for ploychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, 44, 75-92.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, 20(4), 355-371.

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: implications for measurement bias. *Journal of Applied Psychology*, 72(3), 480-483.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: do biased items result in poor measurement? *Educational and Psychological Measurement*, 59(2), 248-269.

Rudner, L., & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. Practical Assessment, Research & Evaluation, 7(26). Retrieved March 3, 2014, from http://files.eric.ed.gov/fulltext/ED458290.pdf

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology*, 67(1), 172-194.

Sackett, P. R., Borneman, M., & Connelly, B. S. (2008). High-stakes testing in education and employment: Evaluating common criticisms regarding validity and fairness. *American Psychologist*, 65, 215-227.

SAS Institute Inc. (2013). *SAS/STAT® 13.1 User's Guide*. Cary, NC: SAS Institute Inc.

Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychomethtrika*, 58(2), 159-194.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, 91, 1291-1306.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. New York: Chapman & Hall/CRC.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, 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.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361-370.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: a DIF analysis of an L2 vocabulary test. *Language Testing*, 17(3), 323-340.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test Validity*, (pp. 147-169). Hillside, NJ: Erlbaum.

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

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

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

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an importance part of instrument development and validation. *Journal of Psychoeducational Assessment*, 29(4), 364-376.

Walker, C. M., Zhang, B., Banks, K., & Cappaert, K. (2012). Establishing effect size guidelines for interpreting the results for differential bundle functioning analyses using SIBTEST. *Educational and Psychological Measurement*, 72(3), 415-434.

Wang, W., & Shih, C. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*, 34(3), 166-180.

Wang, X., Bradlow, E. T., Wainer, H., & Muller, E. S. (2008). A Bayesian method for studying DIF: a cautionary tale filled with surprise and delights. *Journal of Educational and Behavioral Statistics*, 33(3), 363-384.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26(1), 77-87.

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

Wolfinger, R., Tobias, R., & Sall, J. (1994). Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal on Scientific Computing*, 15(6), 1294-1310.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, 33, 42-57.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, 35, 339-361.

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. 337-347). Hillsdale, NJ: Erlbaum.

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

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, 20(2), 136-147.

Appendix A

Table A.1

The ANOVA of Type I Error Rates When Conducting Exploratory Analyses

Source	Sum of Squares	df	Mean Square	F	p	η^2
Intercept	77.476	1	77.476	1572.954	.000	.052
S_DIF	.106	1	.106	2.160	.142	.000
T_DIF	.002	1	.002	.037	.848	.000
S_group	.001	1	.001	.020	.887	.000
T_group	.020	2	.010	.202	.817	.000
Tlvl_NDIF	.322	3	.107	2.176	.089	.000
DIF_lvl	.078	2	.039	.797	.451	.000
S_DIF ×T_DIF	.002	1	.002	.032	.859	.000
S_DIF ×S_group	.000	1	.000	.000	.995	.000
S_DIF ×T_group	.027	2	.014	.278	.757	.000
S_DIF ×Tlvl_NDIF	.000	3	.000	.003	1.000	.000
S_DIF ×DIF_lvl	.039	2	.020	.398	.672	.000
T_DIF ×S_group	.047	1	.047	.950	.330	.000
T_DIF ×T_group	.051	2	.026	.521	.594	.000
T_DIF ×Tlvl_NDIF	.192	3	.064	1.301	.272	.000
T_DIF ×DIF_lvl	.034	2	.017	.350	.705	.000
S_group ×T_group	.060	2	.030	.608	.544	.000
S_group ×Tlvl_NDIF	.170	3	.057	1.150	.327	.000
S_group ×DIF_lvl	.037	2	.018	.371	.690	.000
T_group ×Tlvl_NDIF	.159	6	.026	.537	.780	.000
T_group ×DIF_lvl	.139	4	.035	.705	.589	.000
Tlvl_NDIF ×DIF_lvl	.234	6	.039	.792	.576	.000
S_DIF ×T_DIF ×S_group	.000	1	.000	.005	.944	.000
S_DIF ×T_DIF ×T_group	.793	2	.397	8.052	.000	.001
S_DIF ×T_DIF ×Tlvl_NDIF	.039	3	.013	.265	.851	.000
S_DIF ×T_DIF ×DIF_lvl	.116	2	.058	1.182	.307	.000
S_DIF ×S_group ×T_group	.018	2	.009	.182	.834	.000
S_DIF ×S_group ×Tlvl_NDIF	.083	3	.028	.565	.638	.000
S_DIF ×S_group ×DIF_lvl	.073	2	.036	.741	.477	.000
S_DIF ×T_group ×Tlvl_NDIF	.429	6	.072	1.452	.190	.000
S_DIF ×T_group ×DIF_lvl	.345	4	.086	1.752	.136	.000
S_DIF ×Tlvl_NDIF ×DIF_lvl	.219	6	.037	.741	.616	.000
T_DIF ×S_group ×T_group	.002	2	.001	.022	.978	.000
T_DIF ×S_group ×Tlvl_NDIF	.070	3	.023	.471	.702	.000
T_DIF ×S_group ×DIF_lvl	.033	2	.017	.339	.713	.000
T_DIF ×T_group ×Tlvl_NDIF	.248	6	.041	.838	.540	.000
T_DIF ×T_group ×DIF_lvl	.094	4	.023	.476	.754	.000

T_DIF ×Tlvl_NDIF ×DIF_lvl	.336	6	.056	1.137	.338	.000
S_group ×T_group ×Tlvl_NDIF	.340	6	.057	1.150	.330	.000
S_group ×T_group ×DIF_lvl	.062	4	.015	.312	.870	.000
S_group ×Tlvl_NDIF ×DIF_lvl	.049	6	.008	.166	.986	.000
T_group ×Tlvl_NDIF ×DIF_lvl	.503	12	.042	.851	.597	.000
S_DIF ×T_DIF ×S_group ×T_group	.024	2	.012	.240	.787	.000
S_DIF ×T_DIF ×S_group ×Tlvl_NDIF	.173	3	.058	1.173	.318	.000
S_DIF ×T_DIF ×S_group ×DIF_lvl	.047	2	.024	.479	.619	.000
S_DIF ×T_DIF ×T_group ×Tlvl_NDIF	.383	6	.064	1.297	.254	.000
S_DIF ×T_DIF ×T_group ×DIF_lvl	.123	4	.031	.624	.645	.000
S_DIF ×T_DIF ×Tlvl_NDIF ×DIF_lvl	.263	6	.044	.892	.500	.000
S_DIF ×S_group ×T_group ×Tlvl_NDIF	.493	6	.082	1.670	.124	.000
S_DIF ×S_group ×T_group ×DIF_lvl	.059	4	.015	.299	.878	.000
S_DIF ×S_group ×Tlvl_NDIF ×DIF_lvl	.166	6	.028	.563	.760	.000
S_DIF ×T_group ×Tlvl_NDIF ×DIF_lvl	.480	12	.040	.811	.639	.000
T_DIF ×S_group ×T_group ×Tlvl_NDIF	.052	6	.009	.178	.983	.000
T_DIF ×S_group ×T_group ×DIF_lvl	.223	4	.056	1.134	.338	.000
T_DIF ×S_group ×Tlvl_NDIF ×DIF_lvl	.238	6	.040	.805	.566	.000
T_DIF ×T_group ×Tlvl_NDIF ×DIF_lvl	.489	12	.041	.827	.622	.000
S_group ×T_group ×Tlvl_NDIF ×DIF_lvl	.638	12	.053	1.080	.372	.000
S_DIF ×T_DIF ×S_group ×T_group ×Tlvl_NDIF	.454	6	.076	1.537	.161	.000
S_DIF ×T_DIF ×S_group ×T_group ×DIF_lvl	.179	4	.045	.909	.457	.000
S_DIF ×T_DIF ×S_group ×Tlvl_NDIF ×DIF_lvl	.168	6	.028	.569	.756	.000
S_DIF ×T_DIF ×T_group ×Tlvl_NDIF ×DIF_lvl	.884	12	.074	1.496	.117	.001
S_DIF ×S_group ×T_group ×Tlvl_NDIF ×DIF_lvl	.600	12	.050	1.016	.431	.000
T_DIF ×S_group ×T_group ×Tlvl_NDIF ×DIF_lvl	.554	12	.046	.937	.508	.000
S_DIF ×T_DIF ×S_group ×T_group ×Tlvl_NDIF ×DIF_lvl	.382	12	.032	.646	.805	.000
Error	1402.876	28482	.049			
Total	1493.000	28770				

Note: S_DIF refers the magnitude of student-level DIF; T_DIF refers the magnitude of teacher-level DIF; S_group refers the proportion of students in each student-level manifest group; T_group refers the proportion of teachers in each teacher-level manifest group; Tlvl_NDIF refers the number of teacher-level DIF items; DIF_lvl refers the location of DIF.

Table A.2

The ANOVA of Type I Error Rates When Conducting Confirmatory Analyses

Source	Sum of Squares	df	Mean Square	F	p	η^2
Intercept	23.892	1	23.892	553.696	.000	.044
S_DIF	.093	1	.093	2.156	.142	.000
T_DIF	.008	1	.008	.178	.673	.000
S_group	.048	1	.048	1.103	.294	.000
T_group	.003	2	.002	.036	.965	.000
Tlvl_NDIF	.317	3	.106	2.450	.062	.001
S_DIF ×T_DIF	.047	1	.047	1.099	.295	.000
S_DIF ×S_group	.122	1	.122	2.824	.093	.000
S_DIF ×T_group	.161	2	.081	1.871	.154	.000
S_DIF ×Tlvl_NDIF	.015	3	.005	.120	.949	.000
T_DIF ×S_group	.000	1	.000	.000	1.000	.000
T_DIF ×T_group	.012	2	.006	.134	.874	.000
T_DIF ×Tlvl_NDIF	.204	3	.068	1.573	.194	.000
S_group ×T_group	.000	2	.000	.003	.997	.000
S_group ×Tlvl_NDIF	.086	3	.029	.665	.573	.000
T_group ×Tlvl_NDIF	.406	6	.068	1.569	.152	.001
S_DIF ×T_DIF ×S_group	.002	1	.002	.044	.834	.000
S_DIF ×T_DIF ×T_group	.081	2	.041	.944	.389	.000
S_DIF ×T_DIF ×Tlvl_NDIF	.029	3	.010	.222	.881	.000
S_DIF ×S_group ×T_group	.042	2	.021	.485	.616	.000
S_DIF ×S_group ×Tlvl_NDIF	.089	3	.030	.689	.559	.000
S_DIF ×T_group ×Tlvl_NDIF	.462	6	.077	1.784	.098	.001
T_DIF ×S_group ×T_group	.031	2	.015	.357	.700	.000
T_DIF ×S_group ×Tlvl_NDIF	.169	3	.056	1.309	.269	.000
T_DIF ×T_group ×Tlvl_NDIF	.130	6	.022	.501	.808	.000
S_group ×T_group ×Tlvl_NDIF	.477	6	.080	1.843	.087	.001
S_DIF ×T_DIF ×S_group ×T_group	.022	2	.011	.253	.777	.000
S_DIF ×T_DIF ×S_group ×Tlvl_NDIF	.064	3	.021	.498	.683	.000
S_DIF ×T_DIF ×T_group ×Tlvl_NDIF	.113	6	.019	.435	.856	.000
S_DIF ×S_group ×T_group ×Tlvl_NDIF	.700	6	.117	2.703	.013	.001
T_DIF ×S_group ×T_group ×Tlvl_NDIF	.389	6	.065	1.503	.173	.001
S_DIF ×T_DIF ×S_group ×T_group ×Tlvl_ NDIF	.171	6	.028	.659	.683	.000
Error	513.569	11902	.043			
Total	543.000	11998				

Note: 1. S_DIF refers the magnitude of student-level DIF; T_DIF refers the magnitude of teacher-level DIF; S_group refers the proportion of students in each student-level manifest group; T_group refers the proportion of teachers in each teacher-level manifest group; Tlvl_NDIF refers the number of teacher-level DIF items. 2. The location of DIF is fixed at the teacher level.

Table A.3

Full Factorial ANOVA of Detecting Student-level DIF When Conducting Exploratory Analyses

Source	Sum of Squares	df	Mean Square	F	p	η^2
Intercept	6332.301	1	6332.301	79903.701	.000	.894
S_DIF	207.446	1	207.446	2617.648	.000	**.216**
T_DIF	.011	1	.011	.142	.706	.000
S_group	10.166	1	10.166	128.279	.000	.013
T_group	.072	2	.036	.454	.635	.000
S_DIF × T_DIF	.107	1	.107	1.346	.246	.000
S_DIF × S_group	1.804	1	1.804	22.764	.000	.002
S_DIF × T_group	.001	2	.001	.008	.992	.000
T_DIF × S_group	.020	1	.020	.258	.612	.000
T_DIF × T_group	.589	2	.294	3.714	.024	.001
S_group × T_group	.075	2	.038	.473	.623	.000
S_DIF × T_DIF × S_group	.003	1	.003	.036	.850	.000
S_DIF × T_DIF × T_group	.140	2	.070	.884	.413	.000
S_DIF × S_group × T_group	.332	2	.166	2.092	.123	.000
T_DIF × S_group × T_group	.683	2	.342	4.309	.013	.001
S_DIF × T_DIF × S_group × T_group	.223	2	.112	1.410	.244	.000
Error	753.184	9504	.079			
Total	7307.156	9528				

Note: S_DIF refers the magnitude of student-level DIF; T_DIF refers the magnitude of teacher-level DIF; S_group refers the proportion of students in each student-level manifest group; T_group refers the proportion of teachers in each teacher-level manifest group.

Table A.4

Full Factorial ANOVA of Detecting Teacher-level DIF When Conducting Exploratory Analyses

Source	Sum of Squares	df	Mean Square	F	p	η^2
Intercept	4946.603	1	4946.603	83117.183	.000	.813
S_DIF	.043	1	.043	.721	.396	.000
S_group	.117	1	.117	1.958	.162	.000
T_DIF	532.434	1	532.434	8946.421	.000	**.318**
T_group	286.385	2	143.192	2406.043	.000	**.201**
DIF_lvl	.254	1	.254	4.276	.039	.000
S_DIF ×S_group	.061	1	.061	1.018	.313	.000
S_DIF ×T_DIF	.002	1	.002	.030	.861	.000
S_DIF ×T_group	.014	2	.007	.117	.890	.000
S_DIF ×DIF_lvl	.105	1	.105	1.764	.184	.000
S_group ×T_DIF	.000	1	.000	.001	.969	.000
S_group ×T_group	.024	2	.012	.204	.816	.000
S_group ×DIF_lvl	.223	1	.223	3.755	.053	.000
T_DIF ×T_group	15.399	2	7.700	129.377	.000	.013
T_DIF ×DIF_lvl	.001	1	.001	.015	.903	.000
T_group ×DIF_lvl	.413	2	.207	3.472	.031	.000
S_DIF ×S_group ×T_DIF	.008	1	.008	.141	.707	.000
S_DIF ×S_group ×T_group	.207	2	.104	1.742	.175	.000
S_DIF ×S_group ×DIF_lvl	.225	1	.225	3.778	.052	.000
S_DIF ×T_DIF ×T_group	.366	2	.183	3.079	.046	.000
S_DIF ×T_DIF ×DIF_lvl	.000	1	.000	.000	.988	.000
S_DIF ×T_group ×DIF_lvl	.130	2	.065	1.096	.334	.000
S_group ×T_DIF ×T_group	.096	2	.048	.806	.447	.000
S_group ×T_DIF ×DIF_lvl	.091	1	.091	1.536	.215	.000
S_group ×T_group ×DIF_lvl	.209	2	.104	1.754	.173	.000
T_DIF ×T_group ×DIF_lvl	.072	2	.036	.601	.548	.000
S_DIF ×S_group ×T_DIF ×T_group	.351	2	.176	2.949	.052	.000
S_DIF ×S_group ×T_DIF ×DIF_lvl	.166	1	.166	2.794	.095	.000
S_DIF ×S_group ×T_group ×DIF_lvl	.049	2	.024	.410	.663	.000
S_DIF ×T_DIF ×T_group ×DIF_lvl	.111	2	.056	.935	.393	.000
S_group ×T_DIF ×T_group ×DIF_lvl	.354	2	.177	2.976	.051	.000
S_DIF ×S_group ×T_DIF ×T_group ×DIF_lvl	.072	2	.036	.601	.548	.000
Error	1139.804	19152	.060			
Total	6924.390	19200				

Note: S_DIF refers the magnitude of student-level DIF; T_DIF refers the magnitude of teacher-level DIF; S_group refers the proportion of students in each student-level manifest group; T_group refers the proportion of teachers in each teacher-level manifest group; DIF_lvl refers the location of DIF.

Table A.5

Power of ML-Inter Model when Conducting Exploratory Analyses

Source	Sum of Squares	df	Mean Square	F	Sig.	η^2
Intercept	2938.899	1	2938.899	117753.153	.000	.864
S_DIF	13.120	1	13.120	525.680	.000	.028
T_DIF	167.167	1	167.167	6697.909	.000	**.266**
S_group	.170	1	.170	6.794	.009	.000
T_group	93.611	2	46.805	1875.353	.000	**.169**
DIF_lvl	26.449	1	26.449	1059.733	.000	**.054**
S_DIF ×T_DIF	.233	1	.233	9.318	.002	.001
S_DIF ×S_group	.003	1	.003	.138	.710	.000
S_DIF ×T_group	.056	2	.028	1.115	.328	.000
S_DIF ×DIF_lvl	11.305	1	11.305	452.947	.000	.024
T_DIF ×S_group	.098	1	.098	3.931	.047	.000
T_DIF ×T_group	6.628	2	3.314	132.782	.000	.014
T_DIF ×DIF_lvl	11.297	1	11.297	452.649	.000	.024
S_group ×T_group	1.104	2	.552	22.124	.000	.002
S_group ×DIF_lvl	.479	1	.479	19.180	.000	.001
T_group ×DIF_lvl	12.189	2	6.095	244.197	.000	.026
S_DIF ×T_DIF ×S_group	.030	1	.030	1.192	.275	.000
S_DIF ×T_DIF ×T_group	.404	2	.202	8.099	.000	.001
S_DIF ×T_DIF ×DIF_lvl	.213	1	.213	8.546	.003	.000
S_DIF ×S_group ×T_group	.033	2	.016	.660	.517	.000
S_DIF ×S_group ×DIF_lvl	.031	1	.031	1.251	.263	.000
S_DIF ×T_group ×DIF_lvl	.277	2	.139	5.554	.004	.001
T_DIF ×S_group ×T_group	.025	2	.013	.509	.601	.000
T_DIF ×S_group ×DIF_lvl	.224	1	.224	8.976	.003	.000
T_DIF ×T_group ×DIF_lvl	.501	2	.251	10.042	.000	.001
S_group ×T_group ×DIF_lvl	.176	2	.088	3.516	.030	.000
S_DIF ×T_DIF ×S_group ×T_group	.119	2	.059	2.378	.093	.000
S_DIF ×T_DIF ×S_group ×DIF_lvl	.000	1	.000	.001	.982	.000
S_DIF ×T_DIF ×T_group ×DIF_lvl	.053	2	.026	1.053	.349	.000
S_DIF ×S_group ×T_group ×DIF_lvl	.033	2	.017	.669	.512	.000
T_DIF ×S_group ×T_group ×DIF_lvl	.190	2	.095	3.804	.022	.000
S_DIF ×T_DIF ×S_group ×T_group × DIF_lvl	.022	2	.011	.435	.647	.000
Error	461.626	18496	.025			
Total	3802.907	18544				

Note: S_DIF refers the magnitude of student-level DIF; T_DIF refers the magnitude of teacher-level DIF; S_group refers the proportion of students in each student-level manifest group; T_group refers the proportion of teachers in each teacher-level manifest group; DIF_lvl refers the location of DIF.

Table A.6

Power of ML-Inter Model when DIF at the teacher level when Conducting Exploratory Analyses

Tests of Within-Subjects Contrasts

Source	Sum of Squares	df	Mean Square	F	p	ω^2
DIF_LOC_T	24.010	1	24.010	319.354	.000	.033
DIF_LOC_T ×S_DIF	.003	1	.003	.034	.853	.000
DIF_LOC_T ×T_DIF	2.626	1	2.626	34.921	.000	.004
DIF_LOC_T ×S_group	28.726	1	28.726	382.073	.000	.039
DIF_LOC_T ×T_group	.849	2	.424	5.644	.004	.001
DIF_LOC_T ×S_DIF ×T_DIF	.008	1	.008	.111	.739	.000
DIF_LOC_T ×S_DIF ×S_group	.001	1	.001	.008	.927	.000
DIF_LOC_T ×S_DIF ×T_group	.476	2	.238	3.165	.042	.001
DIF_LOC_T ×T_DIF ×S_group	2.037	1	2.037	27.096	.000	.003
DIF_LOC_T ×T_DIF ×T_group	.788	2	.394	5.238	.005	.001
DIF_LOC_T ×S_group ×T_group	.346	2	.173	2.298	.101	.000
DIF_LOC_T ×S_DIF ×T_DIF ×S_group	.084	1	.084	1.122	.290	.000
DIF_LOC_T ×S_DIF ×T_DIF ×T_group	.059	2	.029	.391	.677	.000
DIF_LOC_T ×S_DIF ×S_group ×T_group	.097	2	.048	.645	.525	.000
DIF_LOC_T ×T_DIF ×S_group ×T_group	.196	2	.098	1.303	.272	.000
DIF_LOC_T ×S_DIF ×T_DIF ×S_group ×T_group	.148	2	.074	.982	.375	.000
Error(DIF_LOC_T)	708.606	9425	.075			

Tests of Between-Subjects Effects

Source	Sum of Squares	df	Mean Square	F	p	ω^2
Intercept	7329.009	1	7329.009	44117.910	.000	.824
S_DIF	.154	1	.154	.926	.336	.000
T_DIF	848.481	1	848.481	5107.541	.000	**.351**
S_group	3.988	1	3.988	24.006	.000	.003
T_group	551.758	2	275.879	1660.689	.000	**.261**
S_DIF ×T_DIF	.004	1	.004	.023	.879	.000
S_DIF ×S_group	.212	1	.212	1.277	.259	.000
S_DIF ×T_group	.369	2	.185	1.111	.329	.000
T_DIF ×S_group	.148	1	.148	.890	.346	.000
T_DIF ×T_group	70.007	2	35.504	213.879	.000	**.062**
S_group ×T_group	1.698	2	.849	5.111	.006	.001
S_DIF ×T_DIF ×S_group	.062	1	.062	.371	.543	.000
S_DIF ×T_DIF ×T_group	.587	2	.293	1.767	.171	.000
S_DIF ×S_group ×T_group	.448	2	.224	1.348	.260	.000
T_DIF ×S_group ×T_group	.589	2	.295	1.773	.170	.000
S_DIF ×T_DIF ×S_group ×T_Sgroup	.373	2	.187	1.124	.325	.000

Error	1565.711	9425	.166

Note: DIF_LOC_T refers the student-teacher manifest groups; S_DIF refers the magnitude of student-level DIF; T_DIF refers the magnitude of teacher-level DIF; S_group refers the proportion of students in each student-level manifest group; T_group refers the proportion of teachers in each teacher-level manifest group.

Table A.7

Power of ML-Inter Model when DIF at both levels when Conducting Exploratory Analyses

Tests of Within-Subjects Effects

Source	Sum of Squares	df	Mean Square	F	p	ω^2
DIF_LOC_B	1931.360	4	482.840	4555.832	.000	**.344**
DIF_LOC_B ×S_DIF	190.557	4	47.639	449.499	.000	**.054**
DIF_LOC_B ×T_DIF	163.339	4	40.835	385.296	.000	**.053**
DIF_LOC_B ×S_group	124.716	4	31.179	294.188	.000	.033
DIF_LOC_B ×T_group	267.405	8	33.426	315.387	.000	**.068**
DIF_LOC_B ×S_DIF ×T_DIF	4.197	4	1.049	9.900	.000	.001
DIF_LOC_B ×S_DIF ×S_group	1.164	4	.291	2.745	.027	.000
DIF_LOC_B ×S_DIF ×T_group	2.561	8	.320	3.021	.002	.001
DIF_LOC_B ×T_DIF ×S_group	2.051	4	.513	4.837	.001	.001
DIF_LOC_B ×T_DIF ×T_group	9.213	8	1.152	10.866	.000	.002
DIF_LOC_B ×S_group ×T_group	3.003	8	.375	3.541	.000	.001
DIF_LOC_B ×S_DIF ×T_DIF ×S_group	.190	4	.047	.448	.774	.000
DIF_LOC_B ×S_DIF ×T_DIF ×T_group	1.509	8	.189	1.779	.076	.000
DIF_LOC_B ×S_DIF ×S_group ×T_group	1.479	8	.185	1.745	.083	.000
DIF_LOC_B ×T_DIF ×S_group ×T_group	1.342	8	.168	1.583	.124	.000
DIF_LOC_B ×S_DIF ×T_DIF ×S_group × T_group	.185	4	.046	.436	.783	.000
Error(DIF_LOC_B)	3676.334	34688	.106			

Tests of Between-Subjects Effects

Source	Sum of Squares	df	Mean Square	F	p	ω^2
Intercept	7851.806	1	7851.806	68812.089	.000	.888
Stu_DIF	105.802	1	105.802	927.232	.000	**.097**
Tea_DIF	196.227	1	196.227	1719.707	.000	**.165**
Ref_group	.319	1	.319	2.798	.094	.000
Tgroup	93.211	2	46.605	408.442	.000	**.086**
Stu_DIF ×Tea_DIF	2.168	1	2.168	18.999	.000	.002
Stu_DIF ×Ref_group	.004	1	.004	.033	.856	.000
Stu_DIF ×Tgroup	.619	2	.310	2.714	.066	.001
Tea_DIF ×Ref_group	.662	1	.662	5.798	.016	.001
Tea_DIF ×Tgroup	8.119	2	4.060	35.578	.000	.008
Ref_group ×Tgroup	4.939	2	2.469	21.641	.000	.005
Stu_DIF ×Tea_DIF ×Ref_group	.301	1	.301	2.641	.104	.000
Stu_DIF ×Tea_DIF ×Tgroup	2.023	2	1.012	8.866	.000	.002
Stu_DIF ×Ref_group ×Tgroup	.246	2	.123	1.079	.340	.000
Tea_DIF ×Ref_group ×Tgroup	.452	2	.226	1.981	.138	.000
Stu_DIF ×Tea_DIF ×Ref_group ×Tgroup	.013	1	.013	.118	.731	.000

Error	989.519	8672	.114

Note: DIF_LOC_B refers the student-teacher manifest groups; S_DIF refers the magnitude of student-level DIF; T_DIF refers the magnitude of teacher-level DIF; S_group refers the proportion of students in each student-level manifest group; T_group refers the proportion of teachers in each teacher-level manifest group.

Table A.8

Power of ML-Inter Model when DIF at both levels when Conducting Confirmatory Analyses

Tests of Within-Subjects Effects

Source	Sum of Squares	df	Mean Square	F	p	ω^2
DIF_LOC_B	1064.634	4	266.159	14133.604	.000	.856
DIF_LOC_B ×S_DIF	3.081	4	.770	40.908	.000	.017
DIF_LOC_B ×T_DIF	20.566	4	5.142	273.030	.000	**.103**
DIF_LOC_B ×S_group	15.540	4	3.885	206.298	.000	**.080**
DIF_LOC_B ×T_group	27.041	8	3.380	179.493	.000	**.131**
DIF_LOC_B ×S_DIF ×T_DIF	22.427	4	5.607	297.734	.000	**.111**
DIF_LOC_B ×S_DIF ×S_group	.634	4	.159	8.419	.000	.004
DIF_LOC_B ×S_DIF ×T_group	4.054	8	.507	26.910	.000	.022
DIF_LOC_B ×T_DIF ×S_group	3.106	4	.776	41.230	.000	.017
DIF_LOC_B ×T_DIF ×T_group	5.509	8	.689	36.570	.000	.030
DIF_LOC_B ×S_group ×T_group	5.209	8	.651	34.577	.000	.028
DIF_LOC_B ×S_DIF ×T_DIF ×S_group	2.515	4	.629	33.384	.000	.014
DIF_LOC_B ×S_DIF ×T_DIF ×T_group	5.439	8	.680	36.100	.000	.029
DIF_LOC_B ×S_DIF ×S_group ×T_group	.892	8	.111	5.920	.000	.005
DIF_LOC_B ×T_DIF ×S_group ×T_group	.341	8	.043	2.262	.021	.002
DIF_LOC_B ×S_DIF ×T_DIF ×S_group × T_group	.052	8	.006	.342	.950	.000
Error(DIF_LOC_B)	178.976	9504	.019			

Tests of Between-Subjects Effects

Source	Sum of Squares	df	Mean Square	F	p	ω^2
Intercept	7134.526	1	7134.526	317811.019	.000	.993
S_DIF	1.643	1	1.643	73.174	.000	.030
T_DIF	18.502	1	18.502	824.201	.000	**.258**
S_group	.560	1	.560	24.960	.000	.010
T_group	18.344	2	9.172	408.564	.000	**.256**
S_DIF ×T_DIF	7.282	1	7.282	324.363	.000	**.120**
S_DIF ×S_group	.972	1	.972	43.298	.000	.018
S_DIF ×T_group	.224	2	.112	4.998	.007	.004
T_DIF ×S_group	1.855	1	1.855	82.634	.000	.034
T_DIF ×T_group	6.527	2	3.264	145.378	.000	**.109**
S_group ×T_group	2.149	2	1.074	47.860	.000	.039
S_DIF ×T_DIF ×S_group	.359	1	.359	15.975	.000	.007
S_DIF ×T_DIF ×T_group	.596	2	.298	13.266	.000	.011
S_DIF ×S_group ×T_group	.054	2	.027	1.212	.298	.001
T_DIF ×S_group ×T_group	.009	2	.004	.197	.821	.000

S_DIF ×T_DIF ×S_group ×T_group	.164	2	.082	3.651	.026	.003
Error	53.339	2376	.022			

Note: DIF_LOC_B refers the student-teacher manifest groups; S_DIF refers the magnitude of student-level DIF; T_DIF refers the magnitude of teacher-level DIF; S_group refers the proportion of students in each student-level manifest group; T_group refers the proportion of teachers in each teacher-level manifest group.

Appendix B

Data generation sample syntax

```
%LET SDIF=0.5;    *STUDENT-LEVEL DIF;
%LET TDIF=0.5;    *TEACHER-LEVEL DIF;
%LET RF=0.5;      *PERCENTAGE OF REFERENCE GROUP;
%LET FF=0.5;      *PERCENTAGE OF FOCAL GROUP;
%LET NDIF=0.125;  *NUMBER OF DIF ITEMS AT TEACHER LEVEL;
%LET TSD=1.44;    *TEACHER EFEFCTIVENESS CUT OFF POINT;

proc iml;
call randseed(0);
*******************TEACHER LEVEL************************;
/*generate teacher ID*/
h=1:100;
h=shape(h,100,1);
/*generate teaching effectiveness from a standard normal distribution*/
TE=randnormal(100,0,1);
TE=shape(TE,100,1);
/*generate teacher level ability*/
theta_mu=randnormal(100,0,1);
theta_mu=shape(theta_mu,100,1);
/*merge teacher ID, teaching effectiveness, and teacher level ability into one
      matrix*/
teacher=h||TE||theta_mu;
/*order the matrix by teaching effectiveness*/
call sort(teacher,2);
/*grouping teachers based on teaching effectiveness: -1.44SD below the mean,
      +1.44SD above the mean*/
m_te=mean(teacher[,2]);  *mean of TE;
v_te=var(teacher[,2]);   *variance of TE;
sd_te=sqrt(v_te);       *standard devaiation of TE;
/*generate teacher level group indicator based on the standard above*/
t_group=j(100,1);
 do i=1 to 100;
   if teacher[i,2] >= &TSD*sd_te+m_te then t_group[i]= 1;
   else if teacher[i,2] <= -&TSD*sd_te+m_te then t_group[i]= 2;
   else t_group[i]=3;
 end;
teacher_new=teacher||t_group;
names={'TID' 'Teff' 'Theta_mu' 'T_group' };
create teacher from teacher_new [colname=names];
append from teacher_new;

*****************************STUDENT LEVEL************************;
theta=randnormal(30,t(teacher_new[,3]),I(100));
theta_stu=shape(t(theta),100*30,1);
TE=repeat(t_group,1,30);
TE1=shape(TE,100*30,1);

u=unique(t_group);
theta_eff=theta_stu[loc(TE1=u[1])];
theta_noneff=theta_stu[loc(TE1=u[2])];
theta_avg=theta_stu[loc(TE1=u[3])];

s1=nrow(theta_eff);
s2=nrow(theta_noneff);
s3=nrow(theta_avg);

/*generate response data of average teachers*/
theta_stu_ref_avg=theta_avg[1:s3*&RF,];
```

```
theta_stu_foc_avg=theta_avg[s3*&RF+1:s3,];
b=j(40,1);
call randgen(b,"Normal");
/*generate b-dif at the studnet level*/
b_stu_dif_avg=j(40,1);
do i=1 to 35;
 b_stu_dif_avg[i]=b[i];
end;
do i=36 to 40;
 b_stu_dif_avg[i]=b[i]+&SDIF;
end;
/*generate response data for reference group*/
z_ref_avg=j(s3*&RF,40);
do i=1 to s3*&RF;
       do j=1 to 40;
              p =exp(theta_stu_ref_avg[i]-b[j])/(1+exp(theta_stu_ref_avg[i]-
       b[j]));
              u=rand('Uniform');
              if p<u then z_ref_avg[i,j]=0;
              if p>u then z_ref_avg[i,j]=1;
       end;
end;

/*generate response data for focal group*/
z_foc_avg=j(s3*&FF,40);
do i=1 to s3*&FF;
       do j=1 to 40;
              p =exp(theta_stu_foc_avg[i]-
       b_stu_dif_avg[j])/(1+exp(theta_stu_foc_avg[i]-b_stu_dif_avg[j]));
              u=rand('Uniform');
              if p<u then z_foc_avg[i,j]=0;
              if p>u then z_foc_avg[i,j]=1;
       end;
end;

z_avg=z_ref_avg//z_foc_avg;
gender_ref_avg=j(s3*&RF,1,1);
gender_foc_avg=j(s3*&FF,1,0);
gender_avg=gender_ref_avg//gender_foc_avg;


/*generate response data of effective teachers*/
theta_stu_ref_eff=theta_eff[1:s1*&RF,];
theta_stu_foc_eff=theta_eff[s1*&RF+1:s1,];

 b_eff=j(40,1);
do i=1 to (1-&NDIF)*40;
 b_eff[i]=b[i];
end;
do i=(1-&NDIF)*40+1 to 40;
 b_eff[i]=b[i]-&TDIF;
end;
/*generate b-dif at the studnet level*/
b_stu_dif_eff=j(40,1);
do i=1 to 35;
 b_stu_dif_eff[i]=b_eff[i];
end;
do i=36 to 40;
 b_stu_dif_eff[i]=b_eff[i]+&SDIF;
end;

/*generate response data of reference group*/
z_ref_eff=j(s1*&RF,40);
```

```
do i=1 to s1*&RF;
       do j=1 to 40;
              p =exp(theta_stu_ref_eff[i]-b_eff[j])/(1+exp(theta_stu_ref_eff[i]-
       b_eff[j]));
              u=rand('Uniform');
              if p<u then z_ref_eff[i,j]=0;
              if p>u then z_ref_eff[i,j]=1;
       end;
end;

/*generate response data for focal group*/
z_foc_eff=j(s1*&FF,40);
do i=1 to s1*&FF;
       do j=1 to 40;
              p =exp(theta_stu_foc_eff[i]-
       b_stu_dif_eff[j])/(1+exp(theta_stu_foc_eff[i]-b_stu_dif_eff[j]));
              u=rand('Uniform');
              if p<u then z_foc_eff[i,j]=0;
              if p>u then z_foc_eff[i,j]=1;
       end;
end;

z_eff=z_ref_eff//z_foc_eff;
gender_ref_eff=j(s1*&RF,1,1);
gender_foc_eff=j(s1*&FF,1,0);
gender_eff=gender_ref_eff//gender_foc_eff;

/*generate response data of noneffective teachers*/
theta_stu_ref_noneff=theta_noneff[1:s2*&RF,];
theta_stu_foc_noneff=theta_noneff[s2*&RF+1:s2,];

b_noneff=j(40,1);
do i=1 to (1-&NDIF)*40;
 b_noneff[i]=b[i];
end;
do i=(1-&NDIF)*40+1 to 40;
 b_noneff[i]=b[i]+&TDIF;
end;
/*generate b-dif at the studnet level*/
b_stu_dif_noneff=j(40,1);
do i=1 to 35;
 b_stu_dif_noneff[i]=b_noneff[i];
end;
do i=36 to 40;
 b_stu_dif_noneff[i]=b_noneff[i]+&SDIF;
end;

/*generate response data for reference group*/
z_ref_noneff=j(s2*&RF,40);
do i=1 to s2*&RF;
       do j=1 to 40;
              p =exp(theta_stu_ref_noneff[i]-
       b_noneff[j])/(1+exp(theta_stu_ref_noneff[i]-b_noneff[j]));
              u=rand('Uniform');
              if p<u then z_ref_noneff[i,j]=0;
              if p>u then z_ref_noneff[i,j]=1;
       end;
end;

/*generate response data for focal group*/
z_foc_noneff=j(s2*&FF,40);
do i=1 to s2*&FF;
       do j=1 to 40;
```

```
            p =exp(theta_stu_foc_noneff[i]-
      b_stu_dif_noneff[j])/(1+exp(theta_stu_foc_noneff[i]-
      b_stu_dif_noneff[j]));
            u=rand('Uniform');
            if p<u then z_foc_noneff[i,j]=0;
            if p>u then z_foc_noneff[i,j]=1;
      end;
end;

z_noneff=z_ref_noneff//z_foc_noneff;
gender_ref_noneff=j(s2*&RF,1,1);
gender_foc_noneff=j(s2*&FF,1,0);
gender_noneff=gender_ref_noneff//gender_foc_noneff;


z=z_noneff//z_avg//z_eff;           *complete response data;
response=shape(z,100*30*40,1);

/*generate a sequence indicating the item number*/
m=(-1)*I(40);
item=repeat(m,100*30,1);
/*generate student ID*/
n=1:30;
n1=repeat(t(n),100,1);
n2=repeat(n1,1,40);
sID=shape(n2,100*30*40,1);

/*generate teacher ID*/
h1=repeat(teacher_new[,1],1,30);
h2=shape(h1,100*30,1);
h3=repeat(h2,1,40);
tID=shape(h3,100*30*40,1);

/*generate student level membership indicator*/
gender=gender_noneff//gender_avg//gender_eff;
s_gender1=repeat(gender,1,40);
s_gender2=shape(s_gender1,100*30*40,1);

/*non-effective variable*/
TE2=repeat(TE1,1,40);
TE3=shape(TE2,100*30*40,1);

/*combine all columns to generate final data for analyses*/

y_data=tID||sID||s_gender2||TE3||item||response;
names={'tID' 'sID' 's_gender' 'TE' 'i1' 'i2' 'i3' 'i4' 'i5' 'i6' 'i7' 'i8' 'i9'
       'i10' 'i11' 'i12' 'i13' 'i14' 'i15' 'i16' 'i17' 'i18' 'i19' 'i20'
       'i21' 'i22' 'i23' 'i24' 'i25' 'i26' 'i27' 'i28' 'i29' 'i30' 'i31' 'i32'
       'i33' 'i34' 'i35' 'i36' 'i37' 'i38' 'i39' 'i40' 'response'};

true_b=b||b_eff||b_noneff||b_stu_dif_avg||b_stu_dif_eff||b_stu_dif_noneff;
name1={'b' 'b_eff' 'b_noneff' 'b_stu_dif_avg' 'b_stu_dif_eff'
       'b_stu_dif_noneff'};

create twolvldata from y_data [colname=names];
append from y_data;
create true_b from true_b [colname=name1];
append from true_b;
quit;
```

Appendix C

Sample syntax of PROC GLIMMIX model

```
************************RASCH DIF MODEL*********************************;
proc glimmix data=twolvldata;
       class sID s_gender;
       model response (Event='1')= i1-i40 s_gender*i36 s_gender*i37 s_gender*i38
s_gender*i39 s_gender*i40/ Dist=Binary link=logit solution noint;
       random intercept / subject=sID TYPE=VC;
       ODS OUTPUT ParameterEstimates=Fix_rasch;
       ODS OUTPUT ConvergenceStatus=Con_rasch;
run;

************************HLM: STUDENT LEVEL COVARIATE********************;
proc glimmix data=twolvldata;
       class tID sID s_gender;
       model response (Event='1')= i1-i40 s_gender*i36 s_gender*i37 s_gender*i38
s_gender*i39 s_gender*i40/ Dist=Binary link=logit solution noint;
       random intercept  / subject=tID type=vc;
       random intercept  / subject=sID(tID) type=vc;
       ODS OUTPUT ParameterEstimates=Fix_slvl;
       ODS OUTPUT ConvergenceStatus=Con_slvl;
run;

************************HLM: TEACHER LEVEL COVARIATE********************;
proc glimmix data=twolvldata;
       class tID sID TE;
       model response (Event='1')= i1-i40 TE*i36 TE*i37 TE*i38 TE*i39 TE*i40 /
Dist=Binary link=logit solution noint;
       random intercept  / subject=tID type=vc;
       random intercept  / subject=sID(tID) type=vc;
       ODS OUTPUT ParameterEstimates=Fix_tlvl;
       ODS OUTPUT ConvergenceStatus=Con_tlvl;
run;

************************HLM: BOTH LEVEL COVARIATE***********************;
proc glimmix data=twolvldata;
       class tID sID s_gender TE;
       model response (Event='1')= i1-i40 s_gender*i36 s_gender*i37 s_gender*i38
s_gender*i39 s_gender*i40 TE*i36 TE*i37 TE*i38 TE*i39 TE*i40 / Dist=Binary
link=logit solution noint;
       random intercept  / subject=tID type=vc;
       random intercept  / subject=sID(tID) type=vc;
       ODS OUTPUT ParameterEstimates=Fix_twolvl;
       ODS OUTPUT ConvergenceStatus=Con_twolvl;
run;
*********************MIXED MODEL: THREE WAY INTERACTION*****************;
proc glimmix data=twolvldata;
       class tID sID s_gender TE;
       model response (Event='1')= i1-i40 s_gender*i36 s_gender*i37 s_gender*i38
s_gender*i39 s_gender*i40 TE*i36 TE*i37 TE*i38 TE*i39 TE*i40 s_gender*TE*i36
s_gender*TE*i37 s_gender*TE*i38 s_gender*TE*i39 s_gender*TE*i40 / Dist=Binary
link=logit solution noint;
       random intercept  / subject=tID type=vc;
       random intercept  / subject=sID(tID) type=vc;
       ODS OUTPUT ParameterEstimates=Fix_mixed;
       ODS OUTPUT ConvergenceStatus=Con_mixed;
run;
```

CURRICULUM VITAE

EDUCATION

Ph. D.	Educational Psychology, University of Wisconsin Milwaukee Concentrations: Educational Statistics and Measurement Minors: Mathematical Statistics & Biostatistics	2014
M. S.	Psychology, Peking University	2009
B. E.	Electrical Engineering, Taiyuan University of Technology	2004

RESEARCH EXPERIENCE

Research Assistant Consulting Office for Research and Evaluation (CORE) 2009 – present
University of Wisconsin - Milwaukee

- Led a research project to evaluate missing data treatments under IRT framework when the latent trait is non-normal.
- Proposed an integral procedure to detect measurement invariance in multilevel data with a within-level violator.
- Presented an empirical data demonstration in a study on within-level measurement invariance
- Cooperated with scholars outside methodological area on a 40-year longitudinal study of ethnicity and gender in U.S. occupations using census data. Took a lead in pulling data, cleaning data, analyzing data and interpreting results.
- Applied factor analyses in the study on validation of the Behavioral Activation for Depression Scale – Short Form with Spanish-speaking Latinos.

Statistical Consultant Consulting Office for Research and Evaluation (CORE) 2009 – present
University of Wisconsin - Milwaukee

- Evaluation of Milwaukee Mathematics Partnership (MMP; NSF founded; http://www4.uwm.edu/Org/mmp/_partners/evaluation.html)
  - Designed surveys for teachers to evaluate MMP program.
  - Evaluated the effect of MMP program on student mathematics achievement using hierarchical regression models and growth models. Interpreted results and reported to the Director of CORE.
  - Managed MMP database, provided descriptively statistical analyses to data manager and contributed to *Milwaukee Mathematics Partnership Final Report* and *Milwaukee Mathematics Partnership Impact Report.*
- Fostering Opportunities foR Tomorrow's Engineers (FORTE; NSF founded; http://www4.uwm.edu/ceas/future_students/forte.cfm)
  - Created and maintained the database of student enrollments and grades.
  - Designed surveys of the satisfactory on the Summer Bridge Program for students in Engineering department.
  - Provided student retention rate annually to PI.
  - Applied chi-square tests to investigate the association between student GPA and retention.
  - Coordinating bi-weekly meeting with PI in Engineering department, interpreting results, and reporting findings.
- Wisconsin State Personnel Development Grant Evaluation (DPI SPDG Evaluation; WI DPI founded; http://sped.dpi.wi.gov/sped_grt_spdgdisc)
  - Created online surveys for Professional Learning Community (PLC) members and Coaches to evaluate the school; and for PLC members to evaluate Coaches.
  - Collecting data and implementing hierarchical linear models to perform the initial analysis.
  - Attending meeting with DPI officials, interpreting results, presenting findings, and

discussing following-up studies

- Junior Achievement Wisconsin (http://www.jawis.org/junior-achievement-of-wisconsin/local-programs)
  - Created online tests for JA BizTown and JA Finance Park
  - Collecting data monthly and reporting summary of data to the vice president of the program
  - Analyzing data, interpreting results, and reporting findings to the vice president

Research Assistant Peking University 2006 – 2009

- Conducted a study to investigate the effect of Emotion Regulation on job burnout of medical staff after the devastating earthquake in Wenchuan, China.
- Carried out a study to explore the moderation effect of the person-organization fit in the relationship between proactive personality and job performance.
- Demonstrated the construct of Chinese Core Self-Evaluation via an empirical study. Implemented the confirmatory factor analysis to validate the structure of Chinese Core Self-Evaluation.

AWARDS AND HONORS

Research Excellence Award	University of Wisconsin Milwaukee	2014
Graduate Research Presentation Award	University of Wisconsin Milwaukee	2014
Graduate Student Travel Award Scholarship	University of Wisconsin Milwaukee	2010 – 2014
Chancellor Graduate Student Award Scholarship	University of Wisconsin Milwaukee	2009 – 2012
Graduate Assistantship	University of Wisconsin Milwaukee	2009 – 2014
Top Scholarship (full tuition award)	Taiyuan University of Technology	2001 – 2004
Excellent Student Award (for top 3% students)	Taiyuan University of Technology	2001 – 2004

PUBLICATIONS

Byars-Winston, A., Fouad, N., & Wen, Y. (under review). A 40-Year Analysis of Race/Ethnicity and Sex in U. S. Occupations, 1970-2010: Implications for Psychological Research, Practice, and Policy. *Journal of Applied Psychology*.

Kim, E. S., Yoon, M., Wen, Y., Luo, W., & Kwok, O. (accepted). Within-level group factorial invariance in multilevel data: Multilevel factor mixture and multilevel MIMIC models. *Structure Equation Modeling*.

Wen, Y., Zhang, B., & Walker, C. (revising). The impact of missing data on non-normal person trait estimation.

Santos, M. M., Kanter, J. W., & Wen, Y. (under review). Validation of the behavioral activation for depression scale – short form (BADS-SF) with Spanish-speaking Latinos. *Behavior Therapy*.

Wen, Y. & Gan, Y. (2008). Proactive Personality and Job Performance: the moderate effect of Person-Organization fit. *The Chinese Journal of Applied Psychology*, 14(2), 118-128.

Hui, C., Gan, Y. Q. & Wen, Y. (2008). An Empirical Research on the Theoretical Construct of Chinese Core Self-evaluation. *Acta Scientiarum Naturalium Universitatis Pekinesis*, 46(1), 141-146.

RESEARCH IN PROGRESS

Wen, Y., Cappaert, K. J. Missing Not At Radom: A Cause of DIF?

Wen, Y., Luo, W., Kim, E. S., & Kwok, O. Testing measurement non-invariance in multilevel data with a within-level violator.

CONFERENCE

Wen, Y., & Walker, C. M. (April, 2014). DIF Analyses in Multilevel Data. Electronic board paper session presented at the 2014 National Council on Measurement in Education, Philadelphia, PA.

Cappaert, K. J., & Wen, Y. (April, 2014). Missing Not At Random: A Cause of DIF? Graduate student poster session presented at the 2014 National Council on Measurement in Education, Philadelphia, PA.

Wen, Y., & Luo, W. (July, 2013). Testing measurement non-invariance in multilevel data with a within-level violator. Paper session presented at the 2013 International Meeting Psychometric Society, Arnhem, the Netherlands.

Wen, Y., & Walker, C. M. (July, 2013). DIF analysis: A combined approach of multilevel IRT models and multilevel mixture IRT models. Paper session presented at the 2013 International Meeting Psychometric Society, Arnhem, the Netherlands.

Wen, Y., & Zhang, B. (April, 2012). Proposing a new IRT guessing model that adjusts both ability and difficulty. Roundtable session presented at the 2012 Annual Meeting of American Educational Research Association, Vancouver, BC, Canada.

Wen, Y., & Zhang, B. (April, 2011). The impact of missing data on non-normal person trait estimation. Graduate student poster session presented at the 2011 National Council on Measurement in Education, New Orleans, LA.

PROFESSIONAL TRAINING

Diagnostic Measurement: Theory, Methods, and Applications. By Jonathan Templin, & Laine Bradshaw	NCME Annual Meeting	2013
Bayesian Networks in Educational Assessment. By Duanli Yan, Russell Almond, Robert Mislevy & David Williamson	NCME Annual Meeting	2013
Test Equating Methods and Practices By Michael Kolen, & Robert L. Brennan	NCME Annual Meeting	2014

TEACHING EXPERIENCE

Teaching Assistant	Educational Statistical Methods I	University of Wisconsin Milwaukee	2012 – 2013
Teaching Assistant	Advanced Statistics in Psychology	Peking University	2008 – 2009
Teaching Assistant	Basic Statistics in Psychology	Peking University	2007 – 2008

WORK EXPERIENCE

Electrical Engineer	Shanxi Kedian Electric Power Design Co. Ltd.	2004 – 2006

SOFTWARE SKILLS

SAS (proficient), SPSS, R, MULTILOG, BILOG, FACT, WINBUGS

PROFESSIONAL SERVICE

Student member	Brad Hanson Award committee	2010 – 2013