EVALUATE MEASUREMENT INVARIANCE ACROSS MULTIPLE GROUPS: A COMPARISON BETWEEN THE ALIGNMENT OPTIMIZATION AND THE RANDOM ITEM EFFECTS MODEL

by

Lida Lin

Bachelor of Science, Nanjing Normal University, 2009

Master of Arts, University of Nebraska-Lincoln, 2012

Submitted to the Graduate Faculty of

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

School of Education

This dissertation was presented

by

Lida Lin

It was defended on

October 26th, 2020

and approved by

Clement Stone, Professor, Department of Educational Foundations Organizations, and Policy

Lan Yu, Associate Professor, Department of Medicine

Co-chair: Feifei Ye, Assistant Professor, RAND Corporation

Co-chair: Suzanne Lane, Professor, Department of Educational Foundations Organizations, and Policy

EVALUATE MEASUREMENT INVARIANCE ACROSS MULTIPLE GROUPS: A COMPARISON BETWEEN THE ALIGNMENT OPTIMIZATION AND THE RANDOM ITEM EFFECTS MODEL

Lida Lin, PhD

University of Pittsburgh, 2020

Participants in achievement tests or psychometric scales can be naturally divided into various sub-groups such as gender, race, social economic status, school district, etc. In order to make meaningful comparison between groups, each item in the test/scale should measure the same underlying construct for participants came from different groups. The increasing implementation of cross-national assessments have raised the question about how to evaluate measurement invariance across a large number of groups. This study compared two relatively new methods—the CFA alignment optimization and the random item effects model—on evaluating measurement invariance. The impact of following factors on the performance of each method were assessed: the proportion of DIF items, type of group mean ability, number of groups, group size, DIF size, and type of DIF. The simulation study demonstrated that both methods performed well in conditions with large number of groups, while they were significantly different from each other. When group size was large and the group mean abilities were equal, both methods would lead to highly accurate parameter estimates, and highly accurate DIF detection rate can be achieved by the alignment method.

Table of Contents

List of Tables

List of Figures

Preface

I would like to thank my committee members: Dr. Suzanne Lane, Dr. Clement Stone, Dr. Feifei Ye, and Dr. Lan Yu. Everyone says the PhD is a long journey, and this has truly been a journey for me on many aspects. Over the past several years, my life has changed so much, and the world is not the same place anymore. Yet you have always been so supportive and encouraging, just like the first time I met you. Thank Dr. Stone and Dr. Yu, you always have the greatest feedback. Thank Dr. Lane for navigating me through difficult transition time. Thank Dr. Ye for your continuous guidance since I first entered this program. Thank you for dedicating so much of your personal time into this dissertation. I am deeply grateful.

I would like to thank my parents and parents-in-law. Thank you for pushing me, and also comforting me.

Last and most importantly, I would like to thank my husband, Yuzong Liu. I would never see the light at the end of the tunnel without you. You are the best thing happened in my life. This dissertation is dedicated to you.

1.0 Introduction

Participants in achievement assessments or psychometric scales can be naturally divided into various groups such as gender, race, social economic status, etc. A fundamental assumption for making meaningful comparisons between groups is that one item should measure the same underlying construct for participants came from different groups. In recent decades, the increasing implementation of cross-national assessments have raised the importance of measurement invariance across large number of groups. This study will compare the performance of two relatively new methods—the alignment optimization and the random item effects model—on their accuracy and efficiency of evaluating measurement invariance across multiple groups. A simulation study with six influential factors: number of groups, group size, DIF size, type of DIF, proportion of DIF items, and type of group mean abilities, will be conducted to compare those two methods. The performance of each method was analyzed from the aspects of convergence, parameter estimation accuracy, and DIF detection accuracy. This study found that the two methods performed different from each other in all conditions. Group size and type of group mean abilities made the greatest impact on the parameter estimation, while both methods performed equally well across conditions with varying number of groups. Overall, both methods are recommended in empirical studies, but which method to chose depends on the specific research interest.

1.1 Background

The violation of measurement invariance is known as differential item functioning (DIF) in the Item Response Theory (IRT) framework. Measurement non-invariance and DIF are two similar terms that can often times be used interchangeably in empirical studies, but there are still different with each other on some aspects. The strict definition of measurement invariance in the factor analysis framework includes four levels of invariance. From the most liberal to the most conservative, they are configural invariance, metric invariance, scalar invariance, and residual invariance. Empirical studies usually only require scalar invariance or metric invariance, because residual invariance is nearly impossible to achieve. Partial measurement invariance is also acceptable in some empirical studies depending on the specific research questions. In the IRT framework, DIF occurs when participants from different groups with the same ability level respond differently to the same item. In an IRT model, unequal item difficulty parameters will lead to uniform DIF. Unequal item difficulty and item discrimination parameters will lead to non-uniform DIF.

A variety of methods have been proposed to detect DIF between two groups. Popular traditional methods include likelihood ratio (LR) test (Thissen, Steinberg et al. 1986, Thissen, Steinberg et al. 1988), the Mantel-Haenszel (MH) test (Holland and Thayer 1988), the standardized p-DIF test (Dorans and Kulick 1986), and the logistic regression model (Swaminathan and Rogers 1990). Other popular models for DIF detecting are the confirmatory factor analysis (CFA) model and the IRT model. The evaluation of measurement invariance in CFA follows a top-down or bottom-up procedure. In the bottom-up procedure, the same configural model will be fitted in both the reference group and the focal group. Then, metric invariance will be tested by constraining factor loadings to be equal between groups. If the model fit indices (χ^2, CFI, TLI, RMSEA, AIC

and BIC) are acceptable, the metric invariance can be confirmed. In the next step, the scalar invariance can be tested by constraining both factor loadings and intercepts simultaneously. If the full measurement invariance is violated, constraints on factor loadings and intercepts can be free one at a time to identify the non-invariant item. The top-down procedure is similar to the bottom-up one but detects measurement non-invariance from the opposite direction.

While the CFA method detecting the DIF items from the group level, the IRT model is able to estimate item parameters of each item directly. Once the item parameters are obtained, DIF can be detected by comparing IRT models with and without constrains on item parameters using the likelihood ratio test. It can also be detected by other statistical tests on item parameters, such as the Lord's chi-square test. In some studies where statistical tests are not required, areas between item characteristic curves (ICC) can also be used to detect DIF in an intuitive way.

Over the past decades, the importance of DIF detection across large number of groups has been recognized due to the increasing popularity of cross-national assessments. Major cross-national assessments are usually administrated across at least 20 countries or education systems. For example, the Trends in International Mathematics and Science Study (TIMSS) evaluates the mathematics and science achievement of students at 4^{th} grade and 8^{th} grade every five years. It consists of questionnaires for students, for teachers, and for schools. From 1995 to 2015, the TIMSS data has been collected from large number of countries ranging from 26 to 50 with an average of 41 countries. In addition, TIMSS has also been administrated in 18 U.S. states over the years, with the maximum participating state number at 13 in 1999. In 2019, TIMSS for 4^{th} grade students was administrated in 58 countries, while TIMSS for 8^{th} grade students was administrated in 39 countries. The Program for International Student Assessment (PISA) measures students' reading, mathematics, and science abilities across the world every three years. PISA includes

questionnaires for both students and schools, and it focus on students' functional skills such as problem-solving skills. From 2000 to 2015, the average number of participating countries of PISA was 59. The minimum and maximum number of countries was 41 in 2003 and 75 in 2009, respectively. In 2015, PISA was administrated in 72 countries along with 3 U.S. states and jurisdictions. The Progress in International Reading Literacy Study (PIRLS) is a comparative assessment that measures the reading literacy of 4^{th} grade students across the world. It consists of questionnaires for students, for teachers, for schools, and an additional curriculum questionnaire. PIRLS is administrated every five years. From 2001 to 2016, the number of participating countries has increased in each administration, with a minimum participating number of 34 in 2001 and a maximum number of 47 in 2016. In 2016, 14 out of 47 participating countries also took an ePIRLS that includes an innovative online reading test. The adaption of ePIRLS was expected to increase the total number of participating countries in the coming administrations.

The popularity of cross-national assessments raises an essential question: how to assure that the instrument is measuring what it was designed to measure equally across all participants from diverse backgrounds. Traditional DIF detection methods introduced above were designed for evaluating measurement invariance between two groups. When there are more than two groups, those methods can still be applied to detect DIF in a pairwise order. However, traditional DIF detection methods are inefficient, and more importantly, not applicable when the number of groups is large.

In recent years, new DIF detection methods have been proposed in response to the increasing number of cross-national studies. Traditional multiple group CFA method (Jöreskog 1970) treats observed scores as a linear function of the common factor score with group-specific factor loadings and intercepts. The multilevel CFA (Muthén 1994) treats group as random, so it

can compare many groups simultaneously by including all groups in a single measurement model. Muthén and Asparouhov (2013a) proposed the Bayesian approximate measurement invariance test in the Bayesian Structural Equation Modeling (BSEM) framework. It assumes approximate measurement invariance and uses approximate zero constraints rather than exact zero constraints in the measurement model. This method allows for small cross loadings exist, which better represents the real data. In the CFA framework, they also proposed an alignment method to evaluate measurement invariance across a large number of groups (Muthén and Asparouhov 2013). Instead of constraining parameters across groups, the alignment method finds the optimal invariance pattern with the minimal number of non-invariant items from the configural model. Different with the Bayesian approximate measurement invariance test and the alignment method, Fox (2010) proposed the random item effects model in the IRT framework. It assumes item characteristics to be random rather than fixed numbers. This model can be seen as an extension of the two-parameter normal ogive response model with normally distributed random item parameters. In this model, DIF can be detected by estimating the variances of item parameters.

1.2 Research Questions

This study will compare the alignment method with the random item effects models on their abilities of assessing measurement invariance across large number of groups. The purpose of this study is to provide empirical guidelines for future cross-national researches. Therefore, this study will generate the data to mimic the features of real cross-national assessments to answer the proposed research questions. It has been demonstrated that both the alignment method and the random item effects model performed better than traditional methods in terms of the accuracy of

parameter estimates and the accuracy of DIF detection under conditions with large number of groups (De Boeck 2008, Fox and Verhagen 2010, Asparouhov and Muthén 2014, Finch 2016, Kim, Cao et al. 2017). In fact, Asparouhov and Muthén (2014) suggested to use the alignment method as a starting point in all conditions, followed with an informed multiple group CFA to avoid the massive amount of model modification work.

Both the alignment method and the random item effects model are computationally expensive. In empirical studies with large dataset, it is important to select a method that best fits the data and provides the most accurate result with the highest efficiency. Muthén and Asparouhov (2013) proposed some criteria for choosing between the alignment method and the random item effects model in varying scenarios. For instance, the alignment method is preferred when the number of factor indicators is small. This specific influential factor will not be discussed in this study because cross-national assessments generally have a decent number of items for each underlying construct. Some other factors may lead to significant discrepancies in DIF detection and parameter estimation, and therefore this study will compare those two methods from the following aspects.

First, sample size will make a difference. In this study, the total sample size is a product of level-2 sample size (number of groups) and level-1 sample size (group size). The random item effects model is essentially a multilevel model. In general, a minimal level-2 sample size of 20 for multilevel modeling is recommended if only regression coefficients are of interest. If the variance of parameter estimates is also of interest, Maas and Hox (2005) pointed out that a minimal level-2 sample size of 50 is required for unbiased estimates of level-2 standard errors. In terms of the alignment method, it is not recommended for dataset with more than 100 groups because of the computation complexity (Asparouhov and Muthén 2014). Based on past data from PISA, TIMSS,

and PIRLS, the number of participating countries in cross-national assessments ranges from 20 to 80. Therefore, this study will evaluate the performance of the alignment method and the random item effects only within this level-2 sample size range.

Second, the proportion of non-invariant items is important. The fundamental assumption of the alignment method is that only a small proportion of items is non-invariant (Muthén and Asparouhov 2013). Asparouhov and Muthén (2014) found that when having 20% DIF items and 100 individuals within each group, the absolute bias of parameter estimates with alignment method were greater than 0.05. This bias decreased as the proportion of DIF items decreased, and as the group size increased. Kim, Cao et al. (2017) also found that the alignment method is optimal when having 20% non-invariant groups as compared to 40% non-invariant groups. This study will compare the performance of the alignment method to the random item effects model under conditions with small to medium proportion of non-invariant items (20% and 40%).

Third, both the alignment method and the random item effects model can be used to identify uniform DIF and non-uniform DIF. This study will compare their performance on detecting both types of DIF along with different DIF sizes. It is known that items with larger DIF are easier to be detected (Finch, 2016). However, only small and medium levels of DIF will be included in this study because items with large DIF rarely appear in well-developed cross-national assessments.

The last factor this study will evaluate is the type of group mean abilities. It has been confirmed that unequal group mean abilities will decrease the accuracy of DIF detection slightly (Stark, Chernyshenko et al. 2006, De Boeck 2008) and affect the coverage rates significantly (Finch 2016). For cross-national assessments, individual abilities of participants from different countries can be dramatically different from each other. This study will compare the alignment method with the random item effects model in the equal ability condition and in the unequal ability

condition. If there is a difference, then researchers will be able to choose the best method depending on their research questions.

Overall, a simulation study will be conducted to answer two main research questions below while controlling all other factors:

1) Under conditions with varying number of groups, group size, proportion of DIF items, and type of group mean abilities, is there any difference in the accuracy of parameter estimates between the alignment method and the random item effects model?
2) Under conditions with varying DIF type and DIF size, is there any difference in the accuracy of DIF detection between different alignment methods?

1.3 Significance of Study

The alignment method and the random item effects model are two relatively new DIF detection methods that have never been compared before on dataset with large number of groups. The purpose of this study is to provide a guideline for future cross-national studies with regards of evaluating measurement invariance. Both methods have great potential and can be applied in a wide range of complicate situations theoretically. For example, the alignment method can be applied on complex survey data by incorporating a weight variable. The random item effects model can introduce other variables on level-2 to better explain the non-invariance across groups. If there is a discrepancy between the performance of these two methods across different conditions, it is important to know which method is superior, or which method is preferred in specific conditions.

From the practical perspective, both methods are computationally expensive, which makes it even more important to choose the most appropriate method to gain the largest efficiency.

2.0 Review of Literature

The increasing implementation of cross-national assessments have raised the importance of measurement invariance. For example, the Programme for International Student Assessment (PISA) examined students' capabilities in mathematics, science, and reading. It has been administrated in 353 countries since 2000, and it was administrated in 72 countries in 2015. The Trends in International Mathematics and Science Study (TIMSS) is another popular cross-national assessment, and it was administrated in 38 countries in 2015. In order to make meaningful comparison between countries, items in the large cross-national assessments should be invariant across groups. For instance, individuals from different countries with the same level of mathematic skills should score the same in a particular math question no matter of their nationality. DIF occurs when the measurement invariance was violated. A variety of methods have been developed to evaluate measurement invariance and to detect DIF. This chapter will introduce some popular methods, with an emphasis on the alignment method and random item effects model.

2.1 Measurement Invariance and Differential Item Functioning

One important prerequisite for psychometric and achievement tests is measurement invariance/equivalence. Each item in the test should measure the same underlying construct across different groups in the same way. Meaningful group comparison can only be made when the assumption of measurement invariance holds. Violation of measurement invariance is also known as differential item functioning. The measurement invariance is a term from the confirmatory

factor analysis (CFA) framework, and DIF is a term from the item response theory (IRT) framework. The concept of measurement non-invariance and DIF are comparable and interchangeable in many situations, particularly in empirical studies, but they are not strictly identical.

2.1.1 Measurement Invariance in CFA Framework

In CFA, the factor model (Muthén 1984) with continuous manifest variables can be written as

$$y_{ij}^{*} = \nu_j + \lambda_j \eta_{ij} + \varepsilon_{ij}, \qquad (1)$$

where $i(i = 1, \ldots I)$ represent the individuals, $j(j = 1, \ldots J)$ represent the items, ν_j is the intercept, λ_j is the factor loading, $\eta_{ij} \sim N(\alpha, \psi)$ is the latent variable, and $\varepsilon_{ij} \sim N(0, \theta_j)$ is the residual. To identify the model, it is common to set $\eta_{ij} \sim N(0,1)$. When the observed response is dichotomous, the factor model (Equation 1) remains the same, but the observed score y_{ij} will be determined by the latent response y_{ij}^{*} as

$$y_{ij} = \begin{cases} 1, & if\ \tau_{ij,1} < y_{ij}^{*} \\ 0, & if\ y_{ij}^{*} \leq \tau_{ij,1} \end{cases}, \qquad (2)$$

where $\tau_{ij,1}$ is the threshold parameter.

There are four levels of measurement invariance in the CFA framework: configural invariance, metric invariance, scalar invariance, and residual invariance (Steenkamp and Baumgartner 1998). The configural invariance is the most liberal assumption, while the residual invariance is the strictest one. The configural invariance, also known as pattern invariance, means that the same construct has been measured by the same items in both groups. To test the configural invariance, two models with the same construct will be tested simultaneously in both groups, and then model fit indices will be evaluated. The metric invariance means that participants from each

group should respond to the same item in the same way. In the CFA model, the metric invariance refers to the same factor loading between groups. It indicates that the item responses relate to the latent construct in the same way. To test the scalar invariance, the intercepts in the CFA model will be compared between two groups. The residual invariance is the strictest invariance among the four. It requires the error variance of each item to be the same between groups.

From configural to residual invariance, each level of invariance is a prerequisite of the next level. Measurement invariance in empirical studies usually refers to scalar invariance and metric invariance. When the strict scalar invariance is violated, partial scalar invariance can still be achieved by allowing some intercepts to be different between groups. Similarly, partial metric invariance can be achieved by releasing certain constraints on factor loadings. Based on the definition, partial measurement invariance holds as long as two factor loadings and intercepts remain invariant between groups (Byrne, Shavelson et al. 1989). But in applied studies, the purpose of testing partial measurement invariance is to identify non-invariance items and make meaningful comparison on the rest of items. Thus, the amount of non-invariance allowed to establish partial measurement invariance depends on the specific research question. In terms of the residual invariance, it does not need to be tested in applied studies because it is nearly impossible to achieve (Byrne 1994).

2.1.2 DIF in IRT Framework

The IRT model can measure an individual's latent ability by item responses using a mathematical function (van der Linden and Hambleton 2013). The basic IRT model assumes unidimensionality and local independence. It means that the respondent's performance can be explained only by one latent trait, and the individual's response to one item does not affect their

response to another item. In the IRT model, DIF occurs when respondents from different groups with the same ability level respond differently to a specific item. DIF can be detected by comparing the item characteristics between groups independently of individual abilities. The advantage of the IRT method is that the item characteristics of each item can be estimated directly with the IRT model. Moreover, graphs of the item characteristic curve (ICC) can be generated to make interpretation and comparison easier.

In the basic one-parameter (1PL) logistic response model, also known as Rasch model (Rasch 1960), the probability of a correct response can be expressed as

$$P\left(Y_{ij}=1\middle|\theta_i,b_j\right)=\frac{\exp(\theta_i-b_j)}{1+\exp(\theta_i-b_j)}, \qquad (3)$$

in which $i(i=1,\dots I)$ is the number of individuals, $j(j=1,\dots J)$ is the number of items, θ_i is the individual ability, and b_j is the item difficulty parameter.

By adding an item discrimination parameter a_j, the model above can be extended to a two-parameter (2PL) logistic model

$$P\left(Y_{ij}=1\middle|\theta_i,a_j,b_j\right)=\frac{\exp[a_j(\theta_i-b_j)]}{1+\exp[a_j(\theta_i-b_j)]}. \qquad (4)$$

The probability of correct response can also be modeled via a normal ogive model (Embretson and Reise 2013)

$$P\left(Y_{ij}=1\middle|\theta_i,a_j,b_j\right)=\Phi\left[a_j\left(\theta_i-b_j\right)\right]=\int_{-\infty}^{a_j(\theta_i-b_j)}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)dt, \qquad (5)$$

where $\Phi(.)$ represents the cumulative normal distribution function, $\int_{-\infty}^{a_j(\theta_i-b_j)}dt$ is the integral notation for the area in the distribution from $-\infty$ to $a_j\left(\theta_i-b_j\right)$, and π equals to 3.14.

As described above, the probability of a correct response on a particular item is a function of individual ability and item characteristics. For an invariant item, the item parameters should

remain constant across different groups. Therefore, DIF occurs when the probability of a correct response differs for individuals with the same level of ability but from different groups.

There are two types of DIF: uniform and non-uniform. The uniform DIF occurs when the conditional dependence between item responses and group membership remains the same across different levels of the latent construct (Osterlind and Everson 2009). In other words, the difference on the probability of correct responses between the focal group and the reference groups is consistent across the ability continuum. The non-uniform DIF occurs when this conditional dependence differs across levels of latent construct. The non-uniform DIF usually indicates an interaction between the underlying ability and the group membership. In the IRT framework, the uniform DIF is DIF in the item difficulty parameter, and the non-uniform DIF is DIF in both item difficulty and item discrimination parameter. It is easier to distinguish uniform and non-uniform DIF from the item characteristic cure (ICC). When uniform DIF occurs on the item difficulty parameter, the ICCs from the reference group and from the focal group will be parallel (Figure 1 left). When non-uniform DIF occurs on both item difficulty and item discrimination parameters, the two ICCs will be crossed (Figure 1 right).

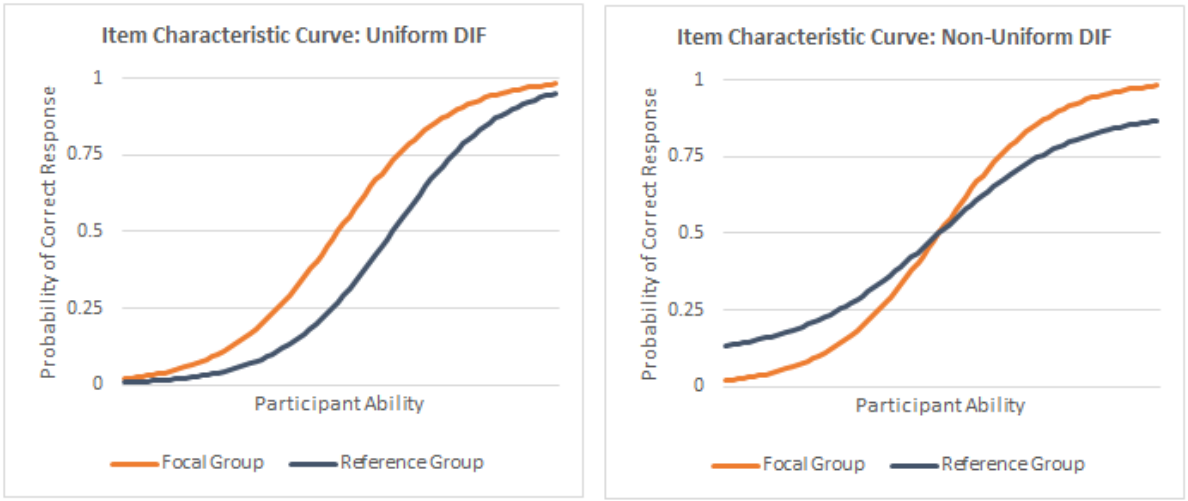

Figure 1. ICC

2.1.3 Link CFA with IRT

In empirical studies, the CFA and IRT are usually seen as alternative methods for each other on DIF detection. Conceptually, the item discrimination parameter in IRT model is analogous to the factor loading in CFA, and the item difficulty parameter is related to the intercept when the mean of the latent construct is fixed to zero (McDonald 2013). Mathematically, the equivalence of the 2PL normal ogive model and the factor analysis model when both using maximum likelihood estimation have been proved by Takane and De Leeuw (1987). This equivalence stands not only for dichotomous responses, but also for ordered categorical responses. Based on Equation (1) and (2), the item parameters in the normal ogive model can be calculated directly from the factor analysis model as:

$$a_j = \left(\frac{\lambda_j}{\sqrt{1-\lambda_j^2}}\right), \qquad (6)$$

$$b_j = \frac{\nu_j}{\lambda_j}. \qquad (7)$$

Although the parameters from CFA and IRT models can be converted to each other in empirical studies, there are three fundamental differences between them. First, the term DIF in the IRT framework does not match the strict form of measurement non-invariance completely. The uniform DIF in the IRT framework matches the scalar non-invariance given metric invariance in the CFA framework, while the non-uniform DIF matches the metric non-invariance. In fact, the concept of DIF and measurement invariance are usually interchangeable in applied studies because: (1) residual invariance is not a concern in empirical studies, and (2) DIF usually refers to uniform DIF by default unless the researcher point it out specifically. The second difference between DIF in the IRT framework and measurement non-invariance in the CFA framework is that the CFA model and the IRT model assess the data on different levels. The IRT approach

explores DIF on the item level, and item characteristics can be directly estimated for DIF analysis. On the other hand, the CFA approach compares model fits on the group level. The overall model fit can be evaluated by CFI, TFI, and RMSEA. Those fit indices can determine if the DIF exists, but they cannot identify which item has DIF. An additional step of comparing chi-square values of two models with and without constraint on a particular item is required to locate the DIF. Third, the IRT model is more appropriate when having dichotomous responses because a logistic regression model is used to link the item responses and the latent ability, while the CFA model is more appropriate when having ordered categorical responses (Raju, Laffitte et al. 2002). Equation (2) can be easily extended to categorical items as

$$\mathrm{y_{ij}} = \begin{cases} C_{\mathrm{ij}}-1, & if\ \tau_{ij,C_{\mathrm{ij}}-1} < y_{ij}^{*} \\ C_{\mathrm{ij}}-2, & if\ \tau_{ij,C_{\mathrm{ij}}-2} < y_{ij}^{*} < \tau_{ij,C_{\mathrm{ij}}-1} \\ \vdots & \\ 1, & if\ \tau_{ij,1} < y_{ij}^{*} \leq \tau_{ij,2} \\ 0, & if\ y_{ij}^{*} \leq \tau_{ij,1} \end{cases}, \qquad (8)$$

where C_{ij} is the number of ordered categories.

Stark, Chernyshenko et al. (2006) proposed a unified DIF detection strategy that can be applied in both CFA framework and IRT framework. This unified strategy was proposed based on the fact that item parameters from the two frameworks are interchangeable. The unified strategy follows three steps. First, fit a free baseline model separately in each framework. Second, specify a baseline model in which the referent items are constrained, and all other items are freely estimated. In the CFA model, loadings of the referent items will be fixed to 1 and intercepts will be constrained to be equal between groups. In the IRT model, the item discriminations and the item difficulties in both groups will be constrained to be equal. Third, various constrained models will be compared to the baseline model. In this step, the loading/item discrimination and intercept/item difficulty will be tested simultaneously using the likelihood ratio test. The chi-

square values from the likelihood ratio test and G^2 from the Bonferroni corrected critical p value will be used to compare models. The CFA approach and the IRT approach are very similar on DIF detection in terms of accuracy and efficiency, though the CFA approach performed better when having polytomous items as compared to having dichotomous items (Stark, Chernyshenko et al. 2006).

2.2 Detect DIF Between Two Groups

In this section, a few traditional DIF detection methods will be introduced first, then the basic CFA and IRT methods will be discussed in detail. These methods were developed for between group DIF detection, but they can be extended to comparisons among a small number of groups following a pairwise sequence.

2.2.1 Traditional DIF Detection Methods

Some popular traditional DIF detection methods based on observed scores includes the IRT likelihood ratio test (Thissen, Steinberg et al. 1986, Thissen, Steinberg et al. 1988), the Mantel-Haenszel (MH) method (Holland and Thayer 1988), the standardized p-DIF test (Dorans and Kulick 1986), and the logistic regression (Swaminathan and Rogers 1990). The IRT likelihood ratio approach takes one item out of the group at a time to build the model, and the remaining items will be treated as the anchor items. Then, each pair of models is compared by the likelihood ratio

test to identify DIF items. The MH method uses the odds of item responses from the focal group and from the reference group to detect DIF. In terms of the standardized p-DIF test, the weighted differences between the focal group and the reference group is evaluated while controlling for underlying individual abilities. The logistic regression approach was proposed as an alternative to the MH method. It incorporates the individual abilities in the regression model to detect uniform DIF based on a null hypothesis that the regression coefficient of the individual ability equals zero. It can also be used to detect non-uniform DIF by incorporating both individual abilities and group membership in the model to evaluate the regression coefficient of the interaction term. The logistic regression approach is more powerful than the MH method for detecting non-uniform DIF (Swaminathan and Rogers 1990).

2.2.2 The CFA Approach

2.2.2.1 Procedure

In order to evaluate measurement invariance, models at different levels of invariance should be tested first, and then model fit indices will be compared to identify where the invariance occurs. In this section, a measurement invariance test between two groups will be described step by step as an example of the CFA procedure.

The first step is to run a model with the same factorial structure, but free factor loadings (except the identification item with the unity factor loading) and intercepts between two groups to test the configural invariance. The configural invariance model is also known as a baseline model in bottom-up measurement invariance studies. Second, constrain the factor loadings to be equal between groups and run a metric invariance model. This is to test whether the two groups have the same meaning of the levels of underlying items. Third, constrain both the factor loadings and

intercepts to be equal to test the scalar invariance. The last step is to constrain residual variances to be equal to test residual invariance. This step is usually skipped in empirical studies since strict measurement invariance rarely exists in real data.

The following model fit indices (Van de Schoot, Lugtig et al. 2012) are commonly used to identify measurement invariance: the chi-square statistic, TLI, CFI, RMSEA, AIC and BIC. Among those, chi-square is the most common indices in applied studies. However, it is sensitive to sample size, so it is less useful when having particularly small or large datasets. The $\Delta\chi^2$ between models can be used to modify the partial invariance model and to identify items with DIF. TLI and CFI can be used to evaluate the baseline model fit. A TLI value larger than 0.95 and a CFI value larger than 0.95 are considered a good fit. RMSEA can be used to examine the closeness of fit. A RMSEA value smaller than 0.05 is good, and the value between 0.05 and 0.08 is acceptable. Although the RMSEA is not sensitive to the sample size, it is sensitive to the model complexity. The AIC and BIC are information indices. They can be used to compare models. The smaller IC value is, the better model fits (Hu and Bentler 1998).

2.2.2.2 Multiple Group CFA

To test measurement invariance across multiple groups, the CFA method follows either a bottom-up or a top-down procedure. The bottom-up method starts with the no invariance model (configural invariance model) and adds invariant item one at a time. The top-down method starts with the full invariance model (scalar invariance model) and release constrained item one at a time. The bottom-up approach is more popular in applied studies, and the test of measurement invariance stops when scalar invariance is met. The top-down approach is more useful when full measurement invariance is required.

Under conditions with more than two groups of participants, the multiple group CFA usually follows the bottom-up procedure. The multiple group CFA is commonly used when having a small number of groups, such as grade, race, SES, etc. A likelihood-ratio test can be performed when adding one invariant item at a time to identify DIF items. When the number of groups is large, three main problems have been found (Asparouhov and Muthén 2014) with multiple group CFA procedure: (1) the possibility of measurement non-invariance is inflated; (2) too many manual steps of model modification are required to achieve measurement invariance or partial measurement invariance, which increase the model complexity; (3) too many model modifications might lead to a completely different scalar model as compared to the configural model. As a result, meaningful group comparison cannot be made across groups.

2.2.3 The IRT Approach

The IRT model is a more straightforward method on DIF detection as compared to the CFA model. After obtaining item parameter estimates and model fit indices, a few methods can be used to detect DIF. All methods discussed in this section can be divided into two categories (Osterlind and Everson 2009): assess DIF as difference in ICC, or assess DIF as difference in item parameters.

2.2.3.1 Assess the Difference on Item Parameters Directly

The simplest method is to compare item parameter estimates (Lord 1980). In a 1PL IRT model, this method is built on a null hypothesis that the item difficulty parameters are equal between the focal group and the reference group. The difference parameter d can be expressed as

$$d = \frac{(\hat{b}_R - \hat{b}_F)}{SE(\hat{b}_R - \hat{b}_F)}, \qquad (9)$$

where $SE(\hat{b}_R - \hat{b}_F)$ is the standard error of difference on item difficulty estimates between the focal group and the reference group, and d follows a standard normal distribution.

Lord (1980) also recommended a chi-square test to examine the differences on item difficulty parameter and item discrimination parameter simultaneously. This method is based on a null hypothesis that all item parameters in both groups are equal. In order to test this null hypothesis, he proposed a chi-square statistic as

$$\chi_L^2 = (v_R - v_F)'(\Sigma_R - \Sigma_F)^{-1}(v_R - v_F), \qquad (10)$$

where v_R is the vector of item parameter estimates for the reference group, v_F is the vector of item parameter estimates for the focal group, Σ_R and Σ_F are the corresponding variance-covariance matrices. The degree of freedom of χ_L^2 is the number of parameters in the model. The Lord's chi-square is a tool to test the equality of item parameters. It can be used to detect both uniform and non-uniform DIF.

2.2.3.2 The Likelihood Ratio Test

The likelihood ratio test compares the likelihood between two models that differ on the item level (Thissen, Steinberg et al. 1988, Camilli and Shepard 1994). It is a test for the overall model fit. One of the two models constrains the item parameter to be equal between the focal group and the reference groups, while the other model allows this parameter to be estimated differently for each group. The likelihood ratios from both groups, written as $L(A)$ and $L(B)$, will be used for a chi-square test

$$G^2 = -2\ln[\frac{L(A)}{L(B)}]. \qquad (11)$$

A significant chi-square test result will indicate a DIF item. Since only one parameter can be constrained at a time, this method is limited in uniform DIF detection on the item difficulty

parameter. When having more than two groups, pairwise comparisons will be made to identify DIF item between each pair of groups.

2.2.3.3 Areas Between ICCs

Raju (1988) first proposed a method to compare both item difficulty and item discrimination parameters based on the area between ICCs in two groups. This method is good for visualizing the differences, but it is not enough to make a statistical inference. Thus, methods such as signed area index (Roju, Van der Linden et al. 1995, Rudner and Gagne 2001) and unsigned area index (Raju 1988), were proposed to measure the areas between ICCs. The signed area index can be written as

$$\text{signed} - \text{area} = \int [P_R(\theta) - P_F(\theta)] d\theta, \qquad (12)$$

where $P_R(\theta)$ is the probability of correct response for the reference group, and $P_F(\theta)$ is the probability of correct response for the focal group. The unsigned area index can be written as

$$\text{unsigned} - \text{area} = \sqrt{\int [P_R(\theta) - P_F(\theta)] d\theta}. \qquad (13)$$

The signed area index can be used to detect uniform DIF, and the unsigned area index can be used to detect non-uniform DIF. The disadvantage of these two methods is that they cannot take the underlying individual abilities into account, so the interpretation of DIF between groups could be misleading to some degree. In addition, this method was built based on the assumption that the guessing parameters are the same between groups, which is not always true in applied studies.

2.3 Detect DIF Across Multiple Groups

All methods mentioned above were proposed for detecting DIF between two groups, the focal group and the reference group. Some of them can be extended to a small number of groups with pairwise comparison, such as the basic CFA method. Some of them can be generalized to a large number of groups, such as the MH method and the Lord's chi-square test. Over the past decades, more methods have been proposed to evaluate DIF among multiple groups, including the alignment method (Muthén and Asparouhov 2013, Asparouhov and Muthén 2014), the random item effects model (Fox 2010), the multilevel CFA (Jak, Oort et al. 2013), the exploratory SEM (Muthén and Asparouhov 2012), and the moderated nonlinear factor analysis (Bauer 2016). This section will focus on the alignment method and the random item effects model, and compare these two approaches from a various of aspects.

2.3.1 The Alignment Method

2.3.1.1 Areas Between ICCs

Section 2.2 has addressed that the CFA top-down or bottom-up method is not practical for detecting DIF across multiple groups. Too many steps of model modification will inflate the model complexity. In addition, the CFA method cannot always guarantee the simplest model with the fewest number of non-invariant parameters. The alignment method was proposed to examine measurement invariance among multiple groups simultaneously (Muthén and Asparouhov 2013, Asparouhov and Muthén 2014, Muthén and Asparouhov 2014). Instead of using the top-down or bottom-up method to test measurement invariance step by step, the alignment method is able to

find the most optimal invariance pattern with the minimal number of non-invariant items. It also provides estimates of factor means, factor variance, intercepts, and loadings that can be used to answer other research questions.

The alignment method includes two steps. The first step is to estimate the configural model (M0) for each group separately. In model M0, all factor means and variances are fixed to 0 and 1, respectively. Factor loadings and intercepts will be freely estimated. In the alignment approach, no additional constraints will be added to model M0. The relationship between M0 and the final model is similar to the relationship between the unrotated model and the rotated model in the EFA framework.

In model M0, the factor mean and variance will be transformed to 0 and 1, respectively. According to

$$\eta_{g0} = (\eta_g - \alpha_g)/\sqrt{\psi_g}. \quad (14)$$

The mean of indicators can be expressed as

$$\nu_{jg,0} = \nu_{jg} + \lambda_{jg}\alpha_g = \nu_{jg} + \frac{\lambda_{jg,0}}{\sqrt{\psi_g}}\alpha_g. \quad (15)$$

The variance of indicators can be expressed as

$$\lambda_{jg,0}^2 = \lambda_{jg}^2\psi_g. \quad (16)$$

Thus,

$$\lambda_{jg,0} = \lambda_{jg}\sqrt{\psi_g}. \quad (17)$$

Every pair of factor mean and variance has a corresponding pair of loading parameters $\lambda_{jg,1}$ and intercept parameters $\nu_{jg,1}$. Based on Equation 17 and 15, the loading and intercept parameters can be calculated from

$$\lambda_{jg,1} = \frac{\lambda_{jg,0}}{\sqrt{\psi_g}}, \quad (18)$$

$$\nu_{jg,1} = \nu_{jg,0} - \frac{\lambda_{jg,0}}{\sqrt{\psi_g}}\alpha_g. \quad (19)$$

The second step is the alignment optimization. The alignment optimization will choose the optimal α_g and ψ_g to minimize the amount of measurement non-invariance. All observed variables are standardized over the population before the optimization, and a simplicity function F was introduced to find the optimal α_g and ψ_g. For every pair of groups, the simplicity function can be written as

$$F = \sum_p \sum_{g_1<g_2} w_{g,g_2}\, f(\lambda_{jg_1,1} - \lambda_{jg_2,1}) + \sum_p \sum_{g_1<g_2} w_{g_1,g_2}\, f(\nu_{jg_1,1} - \nu_{jg_2,1}), \quad (20)$$

where w is the weight factor, and f is a component loss function. The weight factor comes from

$$w_{g,g_2} = \sqrt{N_{g_1} N_{g_2}}, \quad (21)$$

where N is the group size. In the alignment method, the chosen component loss function is

$$f(x) = \sqrt{\sqrt{x^2 + \epsilon}}, \quad (22)$$

where ϵ is a positive small number, so $f(x)$ has a continuous first derivative. In the Mplus default setting, $\epsilon = 0.01$is used. Therefore, $f(x)$ will have no loss when $x = 0$. When $x < 1$, the loss of $f(x)$ will be amplified. When $x > 1$, the loss of $f(x)$ will be attenuated. As a consequence, the simplicity function F will be minimized when there is a large proportion of invariant parameters and a small proportion of non-invariant parameters with large non-invariance. The minimized simplicity function cannot be achieved with a large proportion of medium-sized non-invariant parameters. This leads to a basic assumption of the alignment method that only a small proportion of items can be non-invariant.

All factor means and variances can be obtained from the simplicity function except for the factor mean and variance in the first group. There are two approaches to identify the first group

factor mean and variance: the FIXED alignment optimization and the FREE alignment optimization. In the FIXED alignment approach, α_1 is fixed to zero, and ψ_1 can be calculated from

$$\psi_1 \times ... \times \psi_g = 1. \qquad (23)$$

In the FREE alignment approach, ψ_1 is stil calculated from Equation 23, but α_1 is estimated freely. The the standard errors can be obtained using the delta method.

After obtaining all the parameters, an algorithm (Asparouhov and Muthén 2014) built in Mplus 7 can be used to determine whether a parameter is approximately invariant. This is one of many possible DIF detection methods, but it is the easiest to implement because it can be automatically performed with the "alignment" command in Mplus. This algorithm includes a set of pairwise comparison of item parameters. If the measurement parameter in a particular group is significantly different from the average of that parameter across all groups, it will be considered as a non-invariant parameter. Specifically, this algorithm follows three steps:

1) Conduct a pairwise test for each pair of groups on each parameter. If the *p* value obtained is larger than 0.01, this particular pair of groups will be connected. The largest connected set of groups is considered as the starting set of groups.
2) Calculate the average of each parameter from the starting set of groups. Compare the parameter from each group to the average of the parameters using a significance test. If there is a significant difference, this group will be removed from the invariance set. Otherwise, this group will be added to the invariance set. Note that the *p* value is set to be 0.001 to avoid false positive.
3) Keep repeating Step (2) until a stabilized invariance set has been found. Groups not in the invariance set will be identified to exhibit DIF in that parameter.

The alignment method can not only provide parameter estimates but also label non-invariant items and groups directly in the Mplus output. However, there are two major limitations of the alignment method. First, it assumes that there is only a small proportion of non-invariant items. Second, this method can only be used on the one factor model. The strength and limitations of the alignment method will be further discussed in Section 2.3.3.

2.3.1.2 Estimation Method

Alignment optimization can be estimated by using either the ML method or the Bayes method. Specifically, two Bayesian estimation methods have been proposed: the Bayesian structural equation modeling (BSEM) and the configural method (Muthén and Asparouhov 2012, Asparouhov and Muthén 2014). The difference between those two Bayesian estimations is on model M0.

The configural method requires all factor means and variances in the base model M0 to be fixed to 0 and 1, respectively. All loadings and intercepts in the base model will be freely estimated using non-informative priors, then the posterior distribution of the unaligned configural parameter estimates can be generated. The posterior distribution of the aligned parameter estimates will be obtained by incorporating a simplicity function in each MCMC iteration.

The BSEM method allows factor means and variances from all groups to be freely estimated except for those in the first group. Informative priors with small variances will be used for factor means estimation, so all factor mean estimates will be close to zero. In terms of the first group, the estimates of factor mean and variance depend on the type of alignment. When using the FIXED alignment method, the factor mean will be fixed to 0 and the factor variance will be fixed to 1. When using FREE alignment method, the factor mean in the first group will be estimated freely as in other groups, but the factor variance in the first group will be fixed to 1. Then, all

loadings and intercepts will be estimated using highly correlated priors. As a result, the estimates of loadings and intercepts will be approximately equal across groups. The next step is to generate BSEM loadings and intercepts. The configural loadings and intercepts can be calculated from the BSEM estimates as:

$$\lambda_{jg,0} = \lambda_{jg,1}\sqrt{\psi_g}, \quad (24)$$

$$\nu_{jg,0} = \nu_{jg,1} + \alpha_g \nu_{jg,1}, \quad (25)$$

where $\lambda_{jg,0}$ is the configural loading, $\lambda_{jg,1}$ is the BSEM loading, $\nu_{jg,0}$ is the configural intercept, and $\nu_{jg,1}$ is the BSEM intercpet. The third step is essentially same with the configural method, where aligned configural parameter estimates will be obtained by minimizing the simplicity function in each MCMC iteration.

Both the ML estimation and the Bayesian method can provide asymptotically correct parameter estimates (Asparouhov and Muthén 2014). The ML estimation is computationally less demanding, and it is not affected by prior distributions. On the other hand, the Bayesian estimation generally provides a better model fit than the ML estimation. Asparouhov and Muthén (2014) have demonstrated that the configural method based on non-informative priors leads to slightly more accurate standard errors across a large range of group size. Between the two Bayes methods, the BSEM method is more flexible than the configural method because it allows a small residual covariance existing in the base model (Muthén and Asparouhov 2012). In addition, the BSEM method is preferred when some groups have a relatively large proportion of missing values. The reason is that the BSEM method allows group specific means and variances to be estimated independently, while holding loadings and intercepts approximately equal across groups. In that case, an informative prior obtained from groups with less missing values can be incorporated in the estimation process to stabilize the estimation.

2.3.2 The Random Item Effects Model

The traditional IRT models such as the Rash model or the 2PL normal ogive model treat individuals as random while treating items as fixed. The random item effects model (Fox, 2010) extended the 2PL normal ogive model by treating item characteristics as random. The idea of random items was proposed based on two theoretical reasons (De Boeck 2008): (1) items are drawn from a population; or (2) there is uncertainty existing among item parameters. The random item effects model will fit the data better if there is a random nature of items. Moreover, this model is particularly useful for DIF detection across multiple groups because it allows group specific item characteristics and a common measurement scale to exist at the same time.

2.3.2.1 Model Specification

Based on the Rasch model, De Boeck (2008) proposed two random IRT models to detect DIF between two groups, the random item profile (RIP) model and the random item mixture (RIM) model. The RIP model simply extended the Rasch model (Equation 3) by treating item difficulty parameter as random. De Boeck (2008) suggested putting the estimates of item difficulties from both groups into one robust regression model to detect DIF. The RIP model can be easily extended to multiple groups, and then DIF will be determined by evaluating the variance of item difficulties. The RIM model incorporates a binary latent variable d_j in the RIP model, and the item difficulty parameter can be expressed as

$$b_{jg} = \left(1 - d_j g_p\right)\tilde{b}_{j0} + d_j g_p \tilde{b}_{j1}, \qquad (26)$$

where $g_p = 0$ indicates the reference group, $g_p = 1$ indicates the focal group, and d_j yields a Bernoulli distribution. The item will be classified as a DIF item when the estimate of d_j equals to

1, and be classified as a non-DIF item when the estimate of d_j equals to 0. The RIM model can be considered a special case of the RIP model. The advantage of this model is that it identifies DIF items directly. The disadvantage is that it cannot be extended to multiple groups.

The RIP and RIM model can only be used to detect uniform DIF since only the item difficulty is considered as random. Fox (2010) proposed a random item effects model with both random item difficulty and random item discrimination. This model was built on the 2PL normal ogive model with group-specific item characteristics. When having binary item responses, the model at level-1 can be expressed as

$$P\left(Y_{ijg}=1\middle|\theta_i,\tilde{a}_{jg},\tilde{b}_{jg}\right)=\Phi(\tilde{a}_{jg}\theta_i-\tilde{b}_{jg}), \qquad (27)$$

where $g(g=1,\dots G)$ is the number of items, $\tilde{a}_{F_{jg}}$ and $\tilde{b}_{F_{jg}}$ are the group-specific item discrimination and item difficulty parameters. It should be noticed that this normal ogive model is different from the regular normal ogive model in Equation 5. The item discrimination parameters are the same, but the item difficulty parameter in Equation 27 is the product of item discrimination and item difficulty parameters in Equation 5. Fox (2010) proposed this model so that it can be connected to the CFA more easily. The *a* parameter in the random item effects model can be directly linked to the factor loading in CFA, and the *b* parameter can be directly linked to the intercept.

The group-specific item parameters in the random item effects model are considered to have a common population distribution at a higher level. The level-1 model is cross-classified by the individual abilities and the group-specific item parameters at level-2. The group mean individual abilities and group item parameters will be estimated at the third level. Since this model was built in the multilevel modeling framework, covariates at different levels can be added to

further explain the data. This model can also be applied on datasets with missing values by using item parameter distributions from certain groups as informative priors.

In order to detect DIF, a null hypothesis of cross-group item parameter variances equal to zero will be tested. Measurement invariance (scalar invariance and metric invariance) and partial measurement invariance (partial scalar invariance and partial metric invariance) can be assessed by the discrepancy of different restricted versions of the model.

Based on Equation 27, a model with random individual abilities will be estimated first, known as Model 1. Then the mean and variance of individual abilities will be set to zero and one, respectively. In the next step, a Model 2 with random individual abilities and random item parameters can be estimated. In model M2, the variance of individual ability will be set to one, the sum of intercepts (the product of item difficulty and item discrimination from the 2PL IRT model) in each group will be set to zero, and the product of slopes (item discrimination from the 2PL IRT model) in each group will be set to one. If there is any covariate involved, a Model 3 with covariates on the individual ability parameter will be estimated. Item parameters can be estimated simultaneously by the MCMC algorithm. DIF can be detected by comparing the -2 log-likelihood and DIC of Model 1 and Model 2. It can also be identified by the estimated group variance of item discrimination parameter and item difficulty parameter.

2.3.2.2 Model Estimation

In theory, the random item effects model can be estimated by using either the maximum likelihood (ML) method or the Bayes method. The main problem of the ML estimation is that it requires high-dimensional numerical integration. Several approximations of ML have been proposed to avoid this problem. Instead of the maximum likelihood, the penalized quasi-likelihood (PQL) method (Laird 1978) uses the penalized log-likelihood, which is obtained by a conditional

log likelihood minus the penalty term (Breslow and Clayton 1993). The Laplace approach (Tierney and Kadane 1986) allows random effects within each cluster to be correlated, so that an arbitrary dimension can be built (Raudenbush, Yang et al. 2000). The Laplace approximation can be easily extended to higher orders, so that it can be combined with other estimation methods to improve estimation accuracy. For instance, a PQL method with a fourth order Laplace approximation, known as bias-corrected PQL (Breslow and Lin 1995) can reduce the serious downward bias on parameter estimates. The hierarchical-likelihood (h-likelihood) approximation is a generalization of the Henderson's joint likelihood proposed by Lee and Nelder (1996). It allows the random effects remain invariant across transformations of random components. Bellio and Varin (2005) proposed a pairwise likelihood estimation based on composite likelihood estimation. The pairwise likelihood estimation is a simulation-free estimation method that focuses on pairs of observations, and it can reduce the computation complexity by only computing the bivariate distribution.

All estimation methods mentioned above will lead to biased parameter estimation to some extent, while the Bayesian estimation is generally more accurate than ML approximations. The Bayesian estimation with Monte Carlo Markov chain (MCMC) algorithm (Karim and Zeger 1992) can reduce bias significantly. One additional advantage is that incorporating informative priors in the model can handle dataset with missing values and improve the estimation accuracy. Therefore, Fox and Verhagen (2010) recommended using the Bayesian estimation for the multilevel random item effects model. They calculated the factor means and variances by constraining the sum of intercepts to zero and the product of loadings to one. By treating group-specific item parameters and individual abilities as random, the estimates of item discrimination and item difficulty can be obtained directly. A Bayes factor was incorporated in the estimation process (Fox 2010, Fox and Verhagen 2010, Verhagen and Fox 2013). The Bayes factor was originally defined as

$$\text{BF} = \frac{\frac{P(H_0|\boldsymbol{y})}{P(H_1|\boldsymbol{y})}}{\frac{P(H_0)}{P(H_1)}} = \frac{p(\mathbf{y}|H_0)}{p(\mathbf{y}|H_1)}, \quad (28)$$

where $P(H_0|\boldsymbol{y})$ is the posterior probability of the null hypothesis, $P(H_1|\boldsymbol{y})$ is the posterior probability of the alternative hypothesis, $P(H_0)$ is the prior probability of the null hypothesis, and $P(H_1)$ is the prior probability of the alternative hypothesis (Kass and Raftery 1995). In the random item effects model, the Bayes factor is the ratio of the marginal likelihood of the unrestricted model to marginal likelihood of the restricted model. Since those models are nested, they are considered to have the same conditional distribution. By incorporating nested priors, the Bayes factor can be simplified and be calculated in the unrestricted model. The simplified Bayes factor can be used to test the measurement invariance on item discrimination parameters, threshold parameters, means of latent abilities, and within-group latent variances. The value of the Bayes factor indicates that how much the data is explained by the null hypothesis as opposed to the alternative hypothesis. According to the log_{10} scale (Jeffreys 1998), the null hypothesis will be rejected when the Bayes factor is smaller than ½.

The random item effects model can also be estimated in the SEM framework. From a SEM point of view, the estimation of Fox's model requires the estimation of group specific factor means and factor variances. But it needs to be noted that the group specific factor mean and variance are confounded with group specific intercepts and loadings. Asparouhov and Muthén (2012) proposed another estimation method in a SEM framework. From a two-level SEM perspective, a group specific factor mean $\eta_{2,j}$ can be added to the Equation 1 as

$$\eta_{ij} = \eta_{1,ij} + \eta_{2,j}, \quad (29)$$

where $\eta_{1,ij} \sim N(0,1)$ is at level-1, and $\eta_{2,j} \sim N(0, \psi)$ is at level-2. A group specific indicator $y_{2,j}$ can be added to the factor indicator y_{ij} as

$$y_{ij} = y_{1,ij} + y_{2,j} \quad (30)$$

$$y_{1,ij} = \lambda\eta_{1,ij} + \varepsilon_{ij} \quad (31)$$

$$y_{2,j} = \alpha_j + \lambda\eta_{2,j} \quad (32)$$

$$y_{ij} = \alpha_j + \lambda_j\eta_{1,ij} + \lambda_j\eta_{2,j} + \varepsilon_{ij}, \quad (33)$$

where $\alpha_j \sim N(\alpha, \Sigma)$, Σ is a diagonal matrix, $\lambda_j \sim N(\lambda, \Sigma_2)$, $\eta_{1,ij} \sim N(0, (1+\sigma_j)^2)$, $\varepsilon_{ij} \sim N(0, \Theta)$. According to Equation 32, the group specific factor mean can be identified from the correlation between random intercepts. In addition, $\sigma_j \sim N(0, \sigma)$ is a group specific random parameter, and $(1+\sigma_j)^2$ is a group specific factor variance. Assuming $\eta_{1,ij} \sim N(0,1)$, Equation 33 can then be written as

$$y_{ij} = \alpha_j + \lambda_j(1+\sigma_j)\eta_{1,ij} + \lambda_j\eta_{2,j} + \varepsilon_{ij}. \quad (34)$$

Since $\lambda_j \sim N(\lambda, \Sigma_2)$, λ_j can be written as

$$\lambda_j = \lambda + \lambda_{j0}, \quad (35)$$

where $\lambda_{j0} \sim N(0, \Sigma_2)$. As a result, the random loading in Equation 34 can be written as

$$s_j = (\lambda + \lambda_{j0})(1+\sigma_j) = \lambda + \lambda_{j0} + \lambda\sigma_j + \lambda_{j0}\sigma_j. \quad (36)$$

When majority of items are invariant across groups, both λ_{j0} and σ_j should be very small, which would lead to a negligible $\lambda_{j0}\sigma_j$. As a result, the group specific factor variance can be identified from the correlation between random loadings. In applied studies, a null hypothesis test of $\psi = 0$ examines the factor mean invariance. A null hypothesis test of $\sigma = 0$ examines the factor variance invariance. A null hypothesis test of $\Sigma = 0$ examines the invariance of intercepts. A null hypothesis test of $\Sigma_2 = 0$ examines the invariance of loadings.

Fox (2007) has developed an R package MLIRT using Fortran to estimate their model. But this package has not been updated since Version R 2.15.0 (published in 2012) and no longer

compatible with the latest R. The multilevel SEM (Asparouhov and Muthén 2012) approach is implemented in Mplus Version 7. In this study, the multilevel SEM with Bayesian estimation will be used for data analysis.

2.3.3 Compare the Random Item Effects Model with the Alignment Method

It is known that both methods can be used for evaluating measurement invariance across multiple groups, but the two methods are different from each other from both the conceptual standpoint and the practical standpoint. This section will discuss the similarity and discrepancies between those two methods.

First, both the random item effects model and the alignment method are preferred to the multiple group CFA method under conditions with large number of groups. The multiple group CFA requires a manual model modification process, which may result in too many steps before reaching the acceptable model fit. Because of the large amount of modifications, comparison of factor means across groups could be quite difficult. The alignment method can not only include multiple groups in one model, but also generate parameter estimates for every single item within each group. However, the alignment method is computationally expensive. Asparouhov and Muthén (2014) suggested to use the alignment method as a starting point, followed with an informed multiple group CFA. This approach can avoid the massive amount of model modification work and improve the efficiency of the traditional multiple group CFA method.

In terms of the discrepancy, Muthén and Asparouhov (2013) provided a guideline for choosing between the alignment method and the random item effects model in empirical studies. A decent number of factor indicators are recommended for the random item effects model. However, the sufficient number of factor indicators is sometimes hard to achieve in practical

studies. Asparouhov and Muthén (2014) have demonstrated from a simulation study that the alignment method can still perform well with as few as three factor indicators.

The alignment method is not recommended for studies with more than 100 groups mainly because of the computation complexity. The random item effects model, on the other hand, follows the general rule of multilevel modeling. A minimal higher-level sample size of 20 is suggested for empirical studies if only regression coefficients are of interest. In terms of the variance, Maas and Hox (2005) found that a level-2 sample size of 50 or less will result in biased estimates of the level-2 standard errors.

As mentioned before, the alignment method makes the assumption that only a small proportion of items are non-invariant. It should not be used when that assumption is violated. The random item effects model is preferred when there is little or no prior knowledge about the group characteristics. If researchers already know which groups contribute most to the non-invariance, the alignment method with BSEM estimation will be more efficient. Another advantage of the alignment method is that it does not require a normal distribution of measurement parameters, while the assumption of normality is required for the random item effects model.

In terms of the estimation method, the ML estimation and Bayesian estimation are applicable for both approaches. However, the ML estimation is not suggested for the random item effects model because it requires high-dimensional numerical integration. For the alignment method, both ML and Bayesian estimation can lead to accurate parameter estimates, but the Bayesian estimation is preferred if standard errors are one of the research interests. Generally, the ML estimation is computationally more efficient, and the BSEM estimation is suggested when there is a large proportion of missing values in the dataset.

One of the advantages of the random item effects model is that it can be extended to multiple levels, so that group level or higher level information can be incorporated in the model. However, this model cannot be estimated at this point. Similarly, the alignment method can be applied to complex survey data with multiple sampling strategies and weights in theory, but the estimation method has not been specified yet.

2.3.4 Other Methods

2.3.4.1 Generalized MH Test

The original MH test (Mantel and Haenszel 1959) uses MH chi-square statistic to detect DIF. First, participants will be divided into different levels based on their total scores. For each level, a contingency table of item response and group membership will be constructed. Then, the relationship between item response and group membership will be assessed by the MH statistic across all levels of scores.

When having a small number of groups, the MH test can still be used in a pairwise order, but it leads to an inflation of Type I error. The generalized MH test extended the MH test to multiple groups by allowing different levels of scores to be conditional crossed (Penfield 2001). The generalized MH statistic can be expressed as

$$GMH = (n_j - \mu_j)' V^{-1} (n_j - \mu_j), \qquad (37)$$

where n_j is the vector of item responses from the reference group, μ_j is the vector of expected item responses assuming measurement invariance, and V is the covariance matrix of item responses from the reference group. If the null hypothesis of generalized MH test is rejected, pairwise MH test can be performed to identify the group that causes DIF.

2.3.4.2 The Load's Chi-square Test

The IRT Lord's chi-square test can be extended to multiple groups (Kim, Cohen et al. 1995). In a 2PL model, the chi-square statistic can be written as

$$\chi_L^2 = (Cv)'(C\Sigma_i C')^{-1}(Cv), \qquad (38)$$

where $v = (a_1 b_1, a_2 b_2, \ldots, a_g b_g)$, and C is the contrast matrix of parameters. The contrast matrix needs to be orthogonal. The reference group can be compared to all focal groups simultaneously, and then pairwise comparisons between the reference group and each focal group is applied to identify the problematic group. In order to make item parameters from different groups comparable, item parameters from focal groups will be calibrated to item parameters from the reference group.

2.3.4.3 The Multilevel Confirmatory Factor Analysis

The multilevel CFA extends the basic CFA model to more than two levels by treating groups as random. From the multilevel modeling perspective, the total variance is contributed by both the within-group variance and between-group variance. In the multilevel CFA, a single measurement model takes all groups into consideration. The total variance covariance matrix is contributed by a within group variance covariance matrix and a between group variance covariance matrix. Thus, the measurement model can be written as

$$y_{ij} = y_{W,ij} + y_{B,j} \qquad (39)$$

$$y_{W,ij} = v_W + \lambda_W \eta_{W,ij} + \varepsilon_{W,ij} \qquad (40)$$

$$y_{B,j} = v_B + \lambda_B \eta_{B,j} + \varepsilon_{B,j}, \qquad (41)$$

where $\varepsilon_{W,ij} \sim N(0, \theta_W)$, $\varepsilon_{B,j} \sim N(0, \theta_B)$, and $v_W = 0$.

In terms of testing measurement invariance, the configural invariance is tested in the same way as the basic CFA model. In order to test the metric invariance, a cross-level constraint on factor loadings is added to the model. In other words, all factor loadings of the within group model in Equation 40 are constrained to be equal with factor loadings of the between group model in Equation 41. It is written as $\lambda_W = \lambda_B$. (42)

The scalar invariance is tested differently from the basic CFA model by constraining θ_B to zero (Jak, Oort et al. 2013). The main difference between the multilevel CFA with the random item effects model in multilevel framework is the factor loading. The multilevel CFA assumes equal factor loadings for all groups (Equation 42), but the random item effects model allows factor loading variation.

2.3.4.4 The Bayesian SEM Approach

The Bayesian SEM (BSEM) approach (Muthén and Asparouhov 2012, Muthén and Asparouhov 2013) can be seen as a more flexible form of CFA. It was proposed to take cross-loadings and measurement non-invariance in CFA models into consideration. It assumes only approximate measurement invariance and allows non-invariant items to exist when estimating factor means and variances. In practice, a prior with mean equals to zero and a small variance is commonly used. It generally takes four steps to evaluate the measurement invariance across multiple groups. First, a configural CFA model is estimated using the Bayesian method. All factor means are fixed at 0 and all factor variances are fixed at 1. Second, an approximately invariant prior is incorporated in the model for loadings and intercepts. Third, all factor means are fixed at 0 and factor variances are freely estimated. An optimal prior variance is selected and be compared with the original prior variance. This comparison will determine whether metric invariance should be accepted or be rejected. Finally, only one factor mean is fixed at 0, while the rest of factor

means and all factor variances are freely estimated. Optimal prior variances for factor loadings and intercepts is selected to test scalar invariance. The BSEM method is particularly useful when there is a small proportion of non-invariance items with cross-loadings. When full measurement invariance holds, the original CFA model is more efficient than the BSEM model. When there is a large proportion of non-invariant items, the fundamental assumption of BSEM is violated.

2.3.4.5 The Moderated Nonlinear Factor Analysis

Bauer (2016) proposed a moderated nonlinear factor analysis (MNLFA) which combines the strength of multiple group models and multiple indicator multiple cause (MIMIC) models. The MNLFA model can be used to detect both uniform and non-uniform DIF. Conceptually, the MNLFA approach detects DIF by examining the configural invariance across groups. Technically, DIF in the MNLFA is represented by a form of parameter moderation. The item moderated by the exogenous variable (group indicator) is considered as a DIF item. The MNLFA has two advantages against the multiple group model: (1) the group indicator can be either categorical or continuous; (2) multiple group indicators can be included in the model simultaneously. As comparison, the CFA alignment method can only be applied when having a single underlying construct and a single group indicator. The MNLFA model is now available in Mplus with MIMIC specification and in SAS NLMIXED procedure with specified moderation function.

2.3.4.6 Summary

All methods discussed in this section are relatively new, but this study will only focus on the comparison between the alignment method and the random item effects model, which were built from the two most popular frameworks in researches about measurement invariance. The model specifications and assumptions from the two models are different, but they can be easily

linked to each other. Specifically, no research has been done to compare these two methods in conditions with large number of groups (N>30).

2.4 Previous Simulation Studies

Some studies have been done in recent years to evaluate the performance of the random item effects model and the alignment method on both simulated and empirical datasets. Those two methods have also been compared with other existing DIF detection methods, such as the MH method, the STD p-DIF method, the Lord's chi-square test, and the IRT LR test. This section will introduce these significant studies and summarize six key factors that will influence the performance of DIF detecting methods.

2.4.1 Simulation Studies

2.4.1.1 Studies on the Random Item Effects Model

De Boeck (2008) conducted a simulation study to compare the performance of the RIP model with the MH method, STD p-DIF method, LR test, and logistic regression. They first obtained the item difficulty estimates from the RIP model, then used a robust regression model to detect DIF items. Three factors were controlled in this study: the distribution of item difficulties in the reference group, the pattern of DIF items, and group mean abilities. In each simulation condition, 20 datasets with 20 items were generated using the Rasch model. Two groups with 500 individuals in each group were generated. The distribution of item difficulties was set as either normal or uniform.

There were four DIF items, and the pattern of DIF values was set to be symmetrical (-0.8, 0.8, -1.2, 1.2) and asymmetrical (0.8, 0.8, 1.2, 1.2). The mean abilities in the reference group and in the focal group were set to be equal or unequal. The mean abilities of the reference group and the focal group were fixed at (0,0) in the equal condition, and they were fixed at (0,1) in the unequal condition. The lmer library in R and WinBUGS were used to for estimation. They found that the RIP model with robust regression is generally more accurate than the traditional methods, especially when having asymmetrical DIF pattern. Detecting asymmetrical DIF with traditional methods resulted in large percentage of false positive and false negative.

Fox and Verhagen (2010) conducted a simulation study to evaluate the performance of the random item effects model by assessing the bias of parameter estimates and the convergence properties. They also controlled the distribution of priors to explore the influence of the prior on parameter estimates, which will not be discussed in detail. In this study, 15 dichotomous item responses were generated with 10,000 replications using Fox (2010)'s random item effects model. The number of groups was fixed at 20, and the group size was fixed at 500. Individual abilities, item discriminations, and item difficulties were generated from a two-level model. The final item discriminations ranged from 0.32 to 1.79, and the final item difficulties ranged from -1.16 to 1.32. The MCMC algorithm in MLIRT Package in R was used for parameter estimation. A total of 20,000 iterations with 1,000 burn-in period was used for the MCMC estimation. According to the absolute bias, the mean absolute difference, and the root mean squared differences, he concluded that all item parameters and variances were accurately estimated. He also included five inverse gamma priors (1,1; 0.1, 0.1; 0.01, 0.01; and 1, 0.01) in the Bayesian estimation procedure. As expected, the distribution of priors had no impact on item difficulty estimates and only a small

impact on item discrimination estimates, but it had a significant impact on cross-national item parameter variance estimates.

2.4.1.2 Studies on the Alignment Method

Asparouhov and Muthén (2014) conducted a series of simulation studies to evaluate the quality of the alignment method. They measured the bias of parameter estimates and coverage rate of the alignment method with ML estimation. They also compared the accuracy of ML estimation with the Bayesian estimation. In addition, they compared the performance of the FIXED alignment method with the FREE alignment method across different simulation conditions. Three factors were controlled in this study: number of groups, group size, and degree of non-invariance. In each condition, five items were generated using the single factor CFA model. Number of groups was set at 2, 3, 15, and 60, and group size was set at 100 or 1,000. The percentage of non-invariance was set at 0%, 10%, and 20%. The factor means and variances in Group 1-3 were fixed at (0,1), (0.3,1.5), and (1,1.2). The same distribution pattern kept repeating for the rest of the groups. Similarly, all factor loadings and intercepts for the five items were set at (1,0) except the non-invariant parameters. The non-invariant parameters in Group 1-3 were: $v_5 = 0.5$ and $\lambda_3 = 1.4$, $v_1 = -0.5$ and $\lambda_5 = 0.5$, $v_2 = 0.5$ and $\lambda_4 = 0.3$. The same non-invariant pattern was repeated for the rest of the groups. For the 0% non-invariance condition, all non-invariant item parameters were repeated with the original parameters. For the 10% non-invariance condition, the non- invariant factor loadings in odd groups and non-invariant intercepts in even groups were replaced with original parameters. Mplus 7.11 was used for data generation and parameter estimation. Overall, the FREE alignment method was more accurate than the FIXED alignment method except for the two groups condition. They found that a small overall sample size with large non-invariance led to biased parameter estimates, and bias increased as the size of non-invariance increased. In terms

of the estimation method, the ML estimation tended to overestimate standard errors, while the Bayesian estimation was generally accurate.

Finch (2016) conducted a simulation study to compare the alignment method with the GMH method, generalized logistic regression, and the Lord's chi-square test on detecting uniform DIF across multiple groups. Five factors were controlled in this simulation study: number of groups, reference group sample size, group size ratio, level of DIF, and impact of ability levels. In each simulation condition, 20 dichotomous item responses were generated using the 2PL IRT model with 1,000 replications. There were three levels of number of groups in this study: 2, 3, and 6. There was only one reference group in all conditions, while the rest of groups were focal groups. This study did not take large number of groups into consideration. Even six groups is still a relative small number for large cross-national assessments such as PISA. Three group sizes were used for the reference group: 500, 1,000, and 2,000. The group size of the focal group was chosen to be equal or unequal with the reference group. In the unequal group size conditions, the group size of the focal group was set to be half of the reference group. Since only uniform DIF was examined in this study and the 2PL model was used for data generation, the level of DIF was controlled on the item difficulty parameter at four levels: 0 (no DIF), 0.4 (small DIF), 0.6 (medium DIF), and 0.8 (large DIF). Finally, two types of impacts of individual abilities were included: no impact and with impact. In the no impact condition, the ability levels were equal across groups with a mean of 0. In the with impact condition, one group had an average ability value of 0, while the other groups had average ability values of -0.5. Mplus 7.11 was used for data generation. All methods were compared with the following criteria: Type I error, power, converge rates, estimation bias, standard deviation of parameter estimates, and mean square errors. In the estimation process, Mplus 7.11 was used for the alignment method, and the difR library in R 3.03 was used for the

other methods. They found that the GMH method and alignment method are the most accurate methods in terms of DIF detection. In addition, the alignment method was able to provide unbiased item difficulty estimates in conditions equal group mean abilities. However, large bias on parameter estimates was found in conditions with unequal group mean abilities.

Kim, Cao et al. (2017) conducted a simulation study to compare the alignment method with multiple group CFA, multilevel CFA, multilevel factor mixture modeling, and the Bayesian approximate MI testing across large number of groups. Five factors were controlled in this study: number of groups, group size, percent of non-invairant groups, non-invariance size, and factor mean difference. In each simulation condition, six continuous items were generated with 100 replications using a basic single factor CFA measurement model. SAS 9.4 was used for data generation. There were two levels of number of groups: 25 and 50. Group size was set at three levels: 50, 100, and 1,000. The number of non-invariant items were fixed at two, while the percent of non-invariant groups was set at 20% and 40%. Groups with non-invariant items were considered as a non-invariant cluster. All groups within the non-invariant cluster were generated homogeneously with the same set of factor mean, variance, intercept, and factor loading. Similarly, groups with invariant items were generated homogenously as an invariant cluster. In terms of the non-invariance size, intercepts for five items were generated first using a normal distribution with mean equals to 0 and standard deviation equals to 0.03. The factor loadings of non-invariant items in the non-invariant groups were 0.4 lower than the factor loadings of items in the invariant groups. Similarly, intercepts of non-invariant items in the non-invariant groups were 0.6 higher than items in the invariant groups. Finally, factor means were set at no difference across groups, or 0.5 higher in the non-invariant groups. The performance of those methods was evaluated by the rates of correct measurement invariance detection. In contrast to Asparouhov and Muthén (2014), they

found that the FIXED alignment method performed slightly better than the FREE alignment method. The alignment method was able to identify invariant items and invariant groups in majority of the conditions, but it failed when having 40% non-invariant groups with small group size (GS = 1,000). Overall, the alignment method performed better in small non-invariance conditions than in large non-invariance conditions, which confirmed the terms proposed by Muthén and Asparouhov (2013).

2.4.2 Influential Factors

2.4.2.1 DIF Size

DIF size can be controlled by adding a constant to the original item discrimination and item difficulty in the IRT model, or to the intercept and loading in the CFA model. The accuracy of DIF detection will increase as DIF size increases. The larger the add-in constant is, the easier it will be to correctly identify the DIF items.

Finch (2016) found that when the DIF size is large (0.8 added to item difficulty parameter), a power at or close to 1.0 can be obtained across different number of groups and group sizes. When the DIF size is small (0.4 added to item difficulty parameter) or medium (0.6 added to item difficulty parameter), the power will increase as number of groups and group sizes increases. When having 1,000 individuals or more in each group, a power value of 1.0 can be guaranteed on items with medium DIF size.

2.4.2.2 Number of Groups

The alignment method has a high rate of correctly identifying DIF items across varying number of groups, but it has no significant advantage against traditional methods when the number

of groups is small. Finch (2016) found that both the alignment method and GMH method would led to small Type I error rate (<0.08) when the number of groups was less than or equal to six and the sample size ranged from 500 to 2,000, but the IRT Lord's chi-square test was much more accurate in the same conditions with Type I error rate ranging from 0.01 to 0.02. They also found that the alignment method provided comparable power rate (> 0.8) with GLR and GMH methods when the number of groups is less than or equal to six, while the IRT Lord's chi-square test led to much smaller power rates in the same conditions (Asparouhov and Muthén 2014).

The alignment method is able to provide unbiased parameter estimates across different number of groups. In Finch (2016)'s simulation study with the number of groups ranging from 2 to 6 and group size ranging from 500 to 2,000, the bias of parameter estimates by the alignment method was smaller than 0.02 across all levels of sample size. The standard errors of parameter estimates were smaller than or equal to 0.08, and they decreased as the number of groups increased. The parameter coverage rates were larger than or equal to 0.94, and they increased as the number of groups increased except for items with large DIF values. The mean square errors of parameter estimates were smaller than 0.01 in majority of conditions, except for conditions with six groups and no DIF items (MSE = 0.02). In Asparouhov and Muthén (2014)'s simulation study with number of groups ranging from 2 to 60, unbiased parameter estimates and approximately 95% parameter coverage were found across different number of groups when group size is equal to 1,000. The factor variances were overestimated when having 20% DIF items. But this overestimation decreased as the number of groups increased. From the multilevel modeling perspective, when having a minimum level-2 sample size (20 groups), the random item effects model provided correct parameter estimates and variance estimates with correlations between true and estimated values larger than 0.91 (Fox and Verhagen 2010).

2.4.2.3 Group Size

Generally, a group size of 500 is recommended for a basic 2PL IRT model. In terms of the CFA approach, Stark, Chernyshenko et al. (2006) found that accurate between-group DIF identification can still be achieved even with group size as small as 250. MacCallum, Widaman et al. (1999) conducted a simulation study to investigate the minimum sample size in factor analysis for a good parameter recovery. They found that when having well-determined factors with communalities around 0.5, a sample size ranging from 100 to 200 can lead to accurate parameter estimates. Meade and Lautenschlager (2004) found that a group size of 150 is enough for accurate parameter estimates when having invariant factor loadings between two groups.

The accuracy of DIF detection will increase as the group size increases. When the group size is 1,000 and group number is 2, Stark, Chernyshenko et al. (2006) found that the IRT LR test performed slightly better than the CFA method in terms of the Type I error rate. When having multiple groups and 1,000 individuals in each group, a Type I error rate of 0.05 and a power of 1.0 can be obtained across different number of groups and DIF sizes (Finch 2016). In terms of the random item effects model, De Boeck (2008) found that when having only two groups and 500 individuals in each group, both the RIP model and the MH method identify DIF items correctly with error rates around 3%. The accuracy of those two methods were slightly better than the LR method, and significantly better than the STD p-DIF method. In terms of the random item effects model, correct parameter estimates and variance estimates can be obtained with 500 individuals per group. In terms of the alignment method, Asparouhov and Muthén (2014) found that it can provide unbiased parameter estimates with coverage rate larger than or equal to 95% even with relative small group size (100). However, it cannot correctly detect non-invariant when the group size is 25 (Kim, Cao et al. 2017).

2.4.2.4 Group Mean Abilities

There are two types of group mean abilities: equal and unequal. When having only two groups, Stark, Chernyshenko et al. (2006) found that the unequal group mean abilities will slightly decrease the accuracy of DIF detection made by the CFA method. De Boeck (2008) also found that the error rate of DIF detection made by the RIP model will increase slightly when having unequal group mean abilities as compared to having equal group mean abilities.

Finch (2016) found that the alignment method can provide unbiased parameter estimates across different large number of groups and DIF sizes in the equal group mean ability condition. In conditions where the group mean ability was set to either 0 or -0.5, a bias of approximately 0.5 was found on no DIF, small DIF (0.4 added to the item difficulty parameter), or medium DIF (0.6 added to the item difficulty parameter) items, and an increase in the DIF size would not compromise this bias. Meanwhile, standard deviations of parameter estimates were larger in conditions with unequal group mean abilities than in conditions with equal group mean abilities. They also found that the unequal group mean abilities affected the coverage rates significantly. The convergence rate was around 0.95 in conditions with equal group mean abilities, but it dropped significantly in conditions with unequal group mean abilities.

2.4.2.5 Proportion of DIF

The large proportion of DIF items with small sample size tends to result in biased parameter estimates. Asparouhov and Muthén (2014) found that when having 20% DIF items and 100 individuals within each group, the absolute bias of parameter estimates with alignment method was greater than 0.05. This bias decreased when the proportion of DIF items was 10% and decreased when the group size was 1,000. However, increase the number of groups would not increase the estimation accuracy in such conditions. Kim, Cao et al. (2017) evaluated the

performance of alignment method by controlling the percentage of non-invariant groups instead of DIF items. They found that the alignment method performed better when having 20% non-invariant group than having 40% non-invariant groups.

2.4.2.6 Estimation Method

Fox and Verhagen (2010) conducted a simulation study to investigate the influence of informative priors in the multilevel random item effects model. In their study, the true item discrimination parameter was generated from two levels. The between group item discrimination was generated from a lognormal distribution (1, 0.075), then the group-specific item discrimination was generated from a lognormal distribution with variance equals to 0.02. In the estimation procedure, different inverse gamma (IG) priors were selected for item discrimination parameter. The result showed that informative priors have no impact on item difficulty estimates and little impact on item discrimination estimates. The choice of IG priors tended to influence the cross-group variances of item difficulty parameters slightly. But the extreme informative prior, such as IG (1,1), leaded to consistent large overestimations on cross-group variances.

Asparouhov and Muthén (2014) also compared the Bayesian estimation with non-informative prior to the ML estimation using alignment method. The accuracy of estimation was evaluated by the ratio between average standard errors and the average standard deviations across all replications. They found that the Bayesian estimation provided slightly more accurate standard errors than the ML estimations, though both methods tended to overestimate the standard errors. They also found that the degree of overestimation decreased as group size increased.

3.0 Method

In this study, a MCMC simulation study was conducted to compare the random item effects model with the alignment method in terms of their ability to access measurement invariance across large number of groups. Binary item responses were generated using Fox (2010)'s random item effects model. Six influential factors were controlled in this simulation study: proportion of DIF items, type of group mean abilities, number of groups, group size, DIF size, and types of DIF. Each condition had 100 replications. All datasets were analyzed by the random item effects model, FIXED alignment method, and FREE alignment method, repeatedly. Convergence rate, accuracy of parameter estimation, and accuracy of DIF detection were evaluated to compare the performance of those three methods.

3.1 Simulation factors

A summary of the factor design is presented below (Table 1). In this study, the six influential factors were not fully crossed. One of the design factors was a combination of two influential factors: DIF size and DIF type. Each design factor will be discussed in detail in this section.

Table 1. Simulation Factor Design

Factors	Number of Levels	Factor Design
DIF size × DIF type	4	1. uniform DIF with small DIF value ($a + 0, b + 0.4$) 2. non-uniform DIF with small DIF values ($a + 0.3, b + 0.4$) 3. uniform DIF with medium DIF value ($a + 0, b + 0.6$) 4. non-uniform DIF with medium DIF values ($a + 0.5, b + 0.6$)
Proportion of DIF items	2	1. 20% (4 items) 2. 40% (8 items)
Number of groups	3	1. 24 2. 40 3. 80
Group size	2	1. 100 2. 500
Group mean abilities	2	1. equal: $\theta \sim N(0,1)$ 2. unequal: $\theta \sim N(\beta_g, 1)$, $\beta_g \sim N(0,1)$

3.1.1 DIF Size and Type of DIF

It has been demonstrated that items with relatively large DIF size are easier to be detected even in dataset with small sample size and unequal group mean abilities (Finch 2016). From the practical perspective, items with large DIF size are less likely to be found from a well-developed cross-national assessment. Therefore, this study only explored non-invariant items with small and medium DIF size. In terms of implementation, a constant was added to the particular item parameters to indicate DIF items. Finch (2016) added 0.4, 0.6, 0.8 to item difficulties to represent small, medium, and large DIF. Woods, Cai et al. (2013) added 0.3, 0.5, and 0.7 to both item discriminations and item difficulties to represent small, medium, and large DIF. In this study, the DIF size factor has four levels including both uniform DIF and non-uniform DIF: (1) uniform DIF

with small DIF value ($a + 0$, $b + 0.4$); (2) non-uniform DIF with small DIF values ($a + 0.3$, $b + 0.4$); (3) uniform DIF with medium DIF value ($a + 0$, $b + 0.6$); and (4) non-uniform DIF with medium DIF values ($a + 0.5$, $b + 0.6$). It needs to be noted that the *a* and *b* parameter here indicates the item discrimination parameter and item difficulty parameter in the 2PL IRT model. DIF values were added to *a* and *b* parameter first, then the *a* and *b* parameter were converted to the slope and intercept parameter in Fox's random item effects model for data generation.

3.1.2 Proportion of DIF Items

The total number of items in this study was set at 20. The test length of popular cross-national assessments varies from one to another. PILRS 2011 has 14 items, 16 items, 12 items, and 12 items for each passage, respectively. TIMSS 2011 for Grade 8 mathematics questionnaire has 20 items, 30 items, 20 items, and 18 items for each content domain, respectively. TIMSS 2011 for Grade 8 science questionnaire has 32 items, 18 items, 20 items, and 18 items for each content domain, respectively. Overall, a total of 20 items is close to the average test length of a cross-national assessment, and it is a common setup in IRT simulation studies.

The proportion of DIF items in this study has two levels: 20% (4 items) and 40% (8 items). Muthén and Asparouhov (2014) found that a small number of group (NG = 2 or 3) with a large proportion of non-invariant (20%) items will lead to biased parameter estimates. Kim, Cao et al. (2017) found that the alignment method is more optimal in the 20% non-invariant group conditions than in the 40% non-invariant group conditions. In this study, the proportion of non-invariance was manipulated by the proportion of DIF items. The DIF values were added to only one item in each group.

3.1.3 Number of Groups

The number of groups has three levels: 24, 40, and 80. Since Fox (2010)'s random item effects model was built in a multilevel modeling framework, a minimum level-2 sample size of 20 is required. The smallest number of groups was set to 24 because it can be divided by 8, which is corresponding to the number of DIF items in 40% DIF item conditions. The medium to large number of groups were chosen to reflect the features of real cross-national assessments. From 2000 to 2015, the number of countries participated in PISA ranged from 41 to 75 with a mean of 59. From 1995 to 2015, the number of countries participated in TIMSS ranged from 29 to 50 with a mean of 41. From 2001 to 2016, the number of countries participated in PIRLS ranged from 34 to 47 with a mean of 42. Therefore, 40 groups were chosen to reflect the average number of groups in popular cross-national assessments, and 80 groups were chosen to reflect the maximum number of groups in those studies. Although Muthén and Asparouhov (2013) claimed that the random item effects mdoel will perform better than the alignment method when having more than 100 groups, a group number larger than 80 will not be considered in this study based on the empirical evidence.

3.1.4 Group Size

The group size has two levels: small (100) and large (500). A level-1 sample size of 500 is generally recommended for multilevel 2PL IRT studies. In previous studies, the group size have been ranged from 100 to 2,000 (Muthén and Asparouhov 2014, Finch 2016). The actual group size in real dataset such as PISA or TIMSS can be significantly larger than 500 and often times uneven across groups. Finch (2016) have found that the Type I error rate will drop from .08 to .05 by increasing the group size from 500 to 1,000, but no significant difference on Type I error can be

found when raising the groups size from 1,000 to 2,000. Therefore, a group size of 500, as the smallest group size for a reliable result, was chosen for this study. In addition, a group size of 100 was chosen to evaluate the performance of interested methods with extreme small group size. Uneven group size was not considered in this study, even though it is common in real data.

3.1.5 Group Mean Abilities

The group mean abilities factor has two levels: equal and unequal. Previous studies have been set the group mean abilities in the reference groups and the focal groups to a constant such as (0, 0.5), (0, 1), (0.3, 1.5), etc. (De Boeck 2008, Asparouhov and Muthén 2014, Finch 2016). However, this type of design cannot represent the real data fairly, especially when the number of groups is large. Kim, Cao et al. (2017) randomly generated the factor means from a normal distribution with a mean of 0 and a standard deviation of 0.07, then added a constant of 0.50 to the mean for non-invariant cluster of groups. In this study, the individual abilities were randomly draw from an underlying distribution to better mimic the real cross-national assessments. In equal group mean abilities conditions, all individual abilities were generated from a normal distribution with a mean of 0 and a variance of 1. In unequal group mean abilities conditions, the individual ability parameter θ_{jg} was generated from a normal distribution $\mathrm{N}(\beta_g, 1)$ with a group mean of β_g and a group variance of 1. The group mean ability β_g was generated from a normal distribution with a mean of 0 and a variance of 1 $\mathrm{N}(0, 1)$. As a result, the intraclass correlation (ICC) of the simulated data was approximate 0.5, which is close to the average ICC value (0.48) of the 2015 TIMSS mathematics questionnaire.

3.2 Simulation Procedure

Six factors including number of groups (× 3), group size (× 2), DIF size and DIF type (× 4), proportion of DIF items (× 2) and type of group mean abilities (× 2) were controlled in this study. There were 24 simulation conditions in total with 12 20% DIF items condition and 12 40% DIF items condition. Within each 20% DIF items condition, four items were set up to have four combinations of DIF size and DIF type. Each item was only showing DIF from one group. Within each 40% DIF items condition, eight items were set up to have four combinations of DIF size and DIF type. In other words, each combination in 20% DIF items conditions was repeated twice. Same with the 20% DIF items conditions, each item was only showing DIF from one group. In each simulation condition, 20 dichotomous item responses are generated using SAS 9.4 with 100 replications (Appendix A.1).

Fox (2010)'s random item effects model (Equation 27) was used to generate the probabilities of correct response. The reason of using this model for data generation is that the parameter estimates made by the alignment method and by the random item effects model can be directly compared to the true parameter values. In the equal group mean abilities conditions, individual abilities were generated from a standard normal distribution, $\theta \sim N(0,1)$. In the unequal group mean abilities conditions, individual abilities were generated in two steps as described in Section 3.1.1.5. The item discrimination parameter was generated from a log normal distribution, $lognormal(0, 0.3)$. The item difficulty parameter was generated from a normal distribution with a mean of 0 and variance of 0.25, $normal(0, 0.25)$. The true values for item difficulty and item discrimination parameter were the same for all simulation conditions within each replication, then DIF values were added to the existing item parameters accordingly. One item was only assigned

to have DIF from one group. The item difficulty and item discrimination parameter in IRT model were then converted to the loading and intercept parameter in Fox's random item effects model. The probabilities of correct item response obtained from the random item effects model were compared with a random number generated from a uniform distribution (0, 1). If the probability were larger than the random number, an item response of 1 would be saved. Otherwise, an item response of 0 would be saved.

3.3 Data Analysis

Simulated datasets with binary item responses were analyzed using Mplus Version 8 (Appendix A.2), then results are combined and compared using SAS 9.4 (Appendix A.3).

Each simulated dataset was analyzed using three methods: random item effect model, FIXED alignment method, and FREE alignment method. Since there was no missing value in the simulated dataset, the Bayesian estimation with non-informative prior is used for parameter estimation. When having informative priors, the BSEM estimation can be used for the alignment method, which will lead to unfair comparison between the random item effects model and the alignment method. Two MCMC chains with a minimum number of iterations at 2000 and a maximum number of iterations at 50,000 were used for the Bayesian estimation. The DIF items were identified directly by the alignment method in the Mplus output.

3.4 Outcome Measures

The performance of the alignment method and the random item effects model were evaluated from the following three aspects: convergence, accuracy of parameter recovery, and accuracy of DIF detection.

3.4.1 Convergence

Convergence is essential for a MCMC study. Meaningful inference cannot be made without successful convergence. Mplus determines the convergence of a MCMC algorithm by evaluating the PSR factor for each parameter. By default, Mplus will run two chains simultaneously using multiple processors. It compares the between-chain variation to the sum of between-chain variation and within-chain variation to determine convergence. The MCMC algorithm is considered as converged when the relative between-chain variation is small enough; in other words, when the PSR factor is close to 1. As mentioned above, this study used two processors with a minimum iteration of 2000 and a maximum iteration of 50,000. Mplus treated the first half of the iterations within each chain as burn-in phase. After a minimum of 2000 iteration, the analysis would stop when convergence was reached. If the algorithm failed to converge within 50,000 iterations, Mplus would automatically end with a "NO CONVERGENCE" error message in the output file.

This study evaluated convergence in two steps. In Step 1, any dataset that failed to converge within 50,000 iterations would be automatically counted as non-convergence. Step 2 is to manually screen the converged results by evaluating the trace plots and autocorrelation plots. Trace plots, known as time-series plots, are plots of successive draws (Fox 2010). Slow traversing of the parameter space is an indication of questionable convergence. Figure 2 and Figure 3 are two

examples of trace plots outputted by Mplus. The red vertical line separates the burn-in phase and the rest of iterations. As seen in Figure 2, the second half of iterations shows randomly plotted values around the mean of Markov chain with two chains overlapping in their variation, which indicates a high likelihood of convergence. On the contrast, Figure 3 shows an example of low likelihood of convergence with unstable trends of plotted values. It presents a slow traversing of parameter space within the 10500 iterations. The plotted values were downward, upward, downward, then upward again.

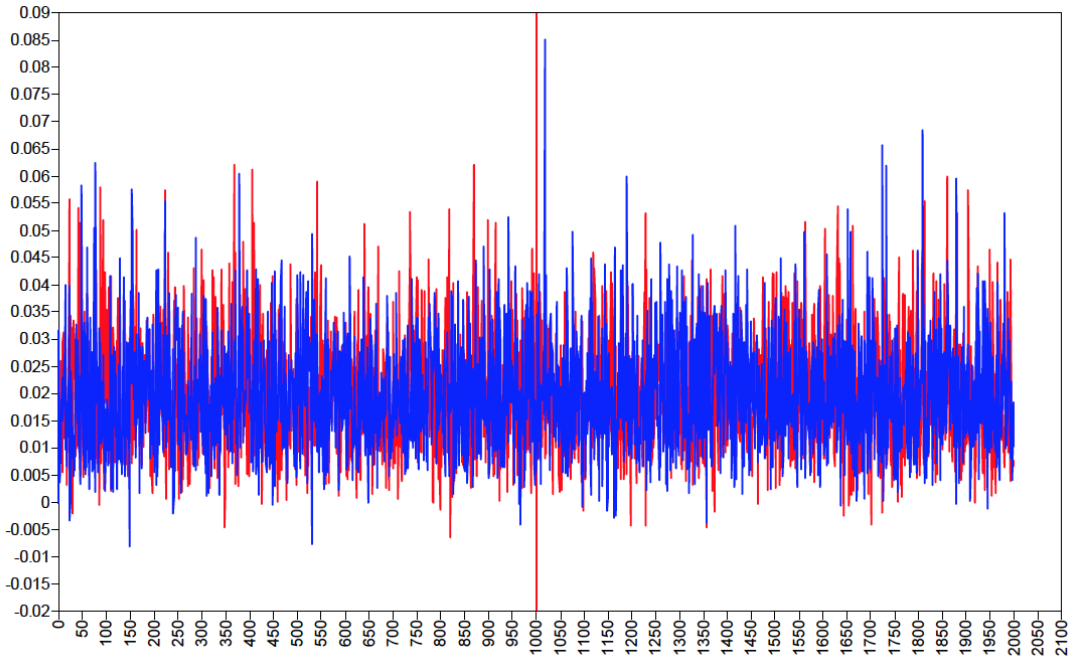

Figure 2. Example Trace Plot from Converged Data

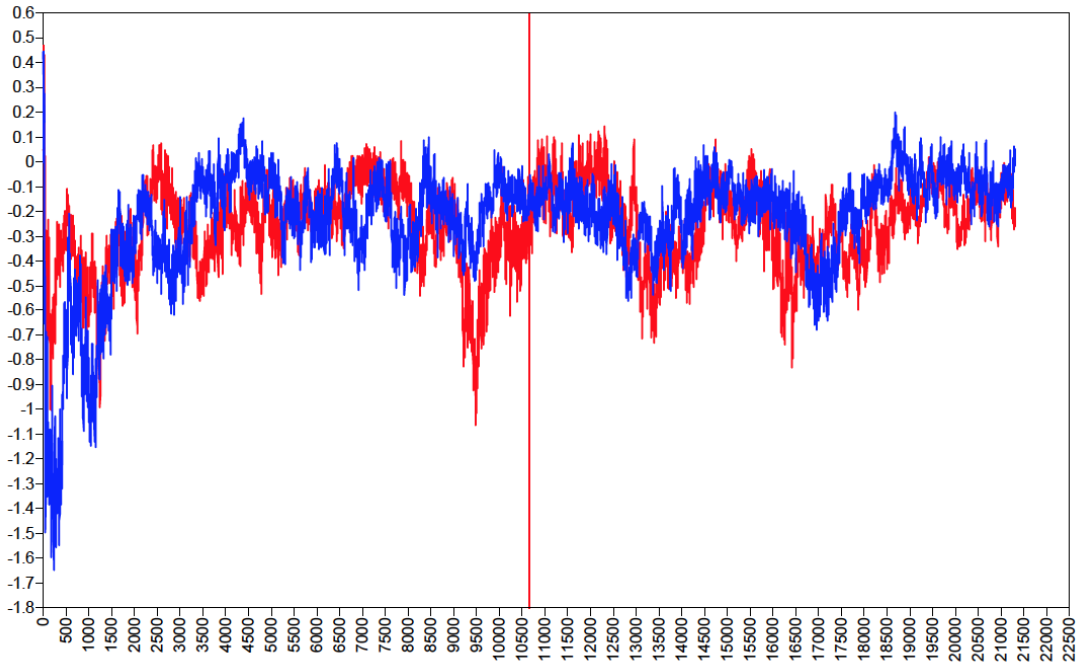

Figure 3. Example Trace Plot from Non-Converged Data

Another tool to check the posterior distribution is the autocorrelation plot, which shows the correlation of parameter values between adjacent iterations for different lags (Fox 2010, Muthén 2010). Smaller value indicates a more efficient Markov chain. Muthen (2010) suggested a value of 0.1 or lower as a cut-off. Figure 4 is an example of the autocorrelation plot generated by Mplus from a converged MCMC algorithm, and Figure 5 is an example of a non-converged algorithm with significant lag. Mplus provides a THIN option for algorithm with high autocorrelation for small lags and decreased autocorrelation with increasing lags, but this option was not adapted in this study. To accelerate the screening process, only plots from datasets that took more than 5,000 iterations to converge were evaluated.

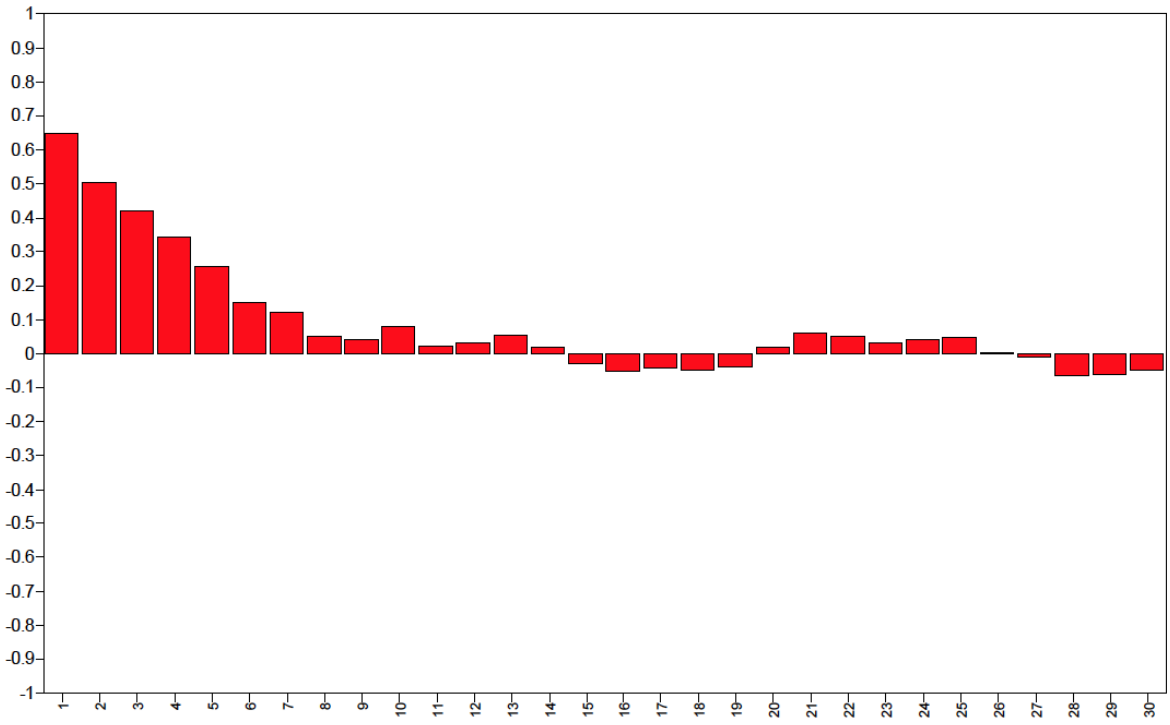

Figure 4. Example of Autocorrelation Plot for Converged Data

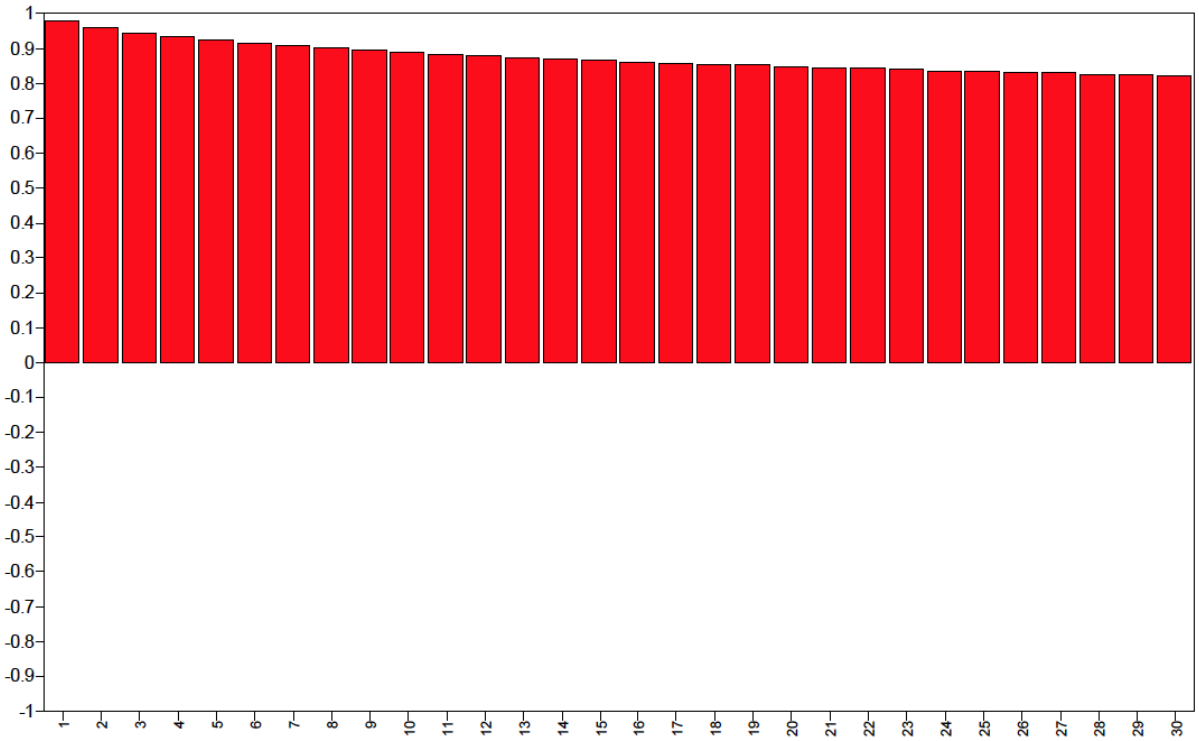

Figure 5. Example of Autocorrelation Plot for Non-Converged Data

3.4.2 Accuracy of Parameter Estimation

The accuracy of parameter estimation was assessed by two criteria: bias and root mean square deviation (RMSD).

Originally, relative bias was considered as a criterion for evaluating estimation accuracy because it makes biases comparable across parameters and scales. However, the relative bias will be amplified when the true parameter value is smaller than 1, and it can be abnormally large as the true parameter value getting close to 0. Preliminary analysis found that when taking the average of relative biases across replications and items, a small proportion of extremely large values would shift the mean dramatically. Therefore, this study chose the true bias over the relative bias as an evaluation criterion. The true bias of a parameter can be expressed as

$$\text{Bias} = \hat{y}_{ij}\text{-}y_{ij}, \qquad (43)$$

where $\hat{y}_{ij}$ is the parameter estimate, and y_{ij} is the true value of this parameter. When y_{ij} represents the slope parameter, $\hat{y}_{ij}$ is the estimate of slope. When y_{ij} represents the intercept parameter, $\hat{y}_{ij}$ is the estimate of intercept with a minus sign. The overall bias that used for analysis was the mean of biases across 20 item parameters. For the random item effects model, the bias can be obtained on the individual level first, then averaged within the group. For the alignment method, parameter estimates were obtained on the group level directly. Mixed ANOVA was used to compare the biases of slope and intercept parameters across simulation conditions.

The accuracy of parameter estimates was also assessed by the RMSD of expected score, slope parameter, and intercept parameter. The RMSD of item parameter can be expressed as

$$RMSD = \sqrt{\frac{\sum(y_{ij}-\hat{y}_{ij})^2}{n}}, \qquad (44)$$

where $\hat{y}_{ij}$ is the parameter estimate, y_{ij} is the true value of this parameter, and *n* is the number of items. When y_{ij} represents the slope parameter, $\hat{y}_{ij}$ is the estimate of slope. When y_{ij} represents the intercept parameter, $\hat{y}_{ij}$ is the estimate of intercept with a minus sign. The RMSD of expected scores was chosen for evaluation because it does not require model parameters to be on the same scale. By calculating the RMSD of expected scores, meaningful comparison can be made across different studies. For an item *j* with dichotomous response, the expected score can be expressed as

$$E(y_j) = 1 \times P(y = 1) + 0 \times P(y = 0) = P(y = 1). \quad (45)$$

In this study, the expected score was the true value $P(Y_{ijg} = 1|\theta_i, \tilde{a}_{jg}, \tilde{b}_{jg})$ generated from Equation 27 in the simulation procedure. As a result, the RMSD of expected score can be calculated using Equation 44 by replacing the true value with the expected score.

3.4.3 Accuracy of DIF Detection

One of the greatest advantages of the alignment method is that it can directly flag DIF items in the Mplus output. The accuracy of this identification was assessed by the power and Type I error rate. The power is the rate of correctly identifying a DIF item, and Type I error is the rate of incorrectly identifying a DIF item while it is not. Power and Type I error rate were calculated for each simulated dataset, then averaged across 100 replications.

4.0 Results

4.1 Data Generation

Dichotomous item responses were generated using Fox (2010)'s random item effects model in SAS 9.4. Six factors were controlled in the simulation process: proportion of DIF items, group mean individual abilities, number of groups, group size, and a combination of DIF type and DIF size. A total of 24 simulation conditions were generated with 150 replications for each condition. The reason why number of replications was increased to 150 will be discussed in Section 4.2. Each replication was analyzed by three methods: random item effect model, FIXED alignment method, and FREE alignment method. Mplus 8 was used for data analysis. The results including convergence information, parameter estimates, and DIF detection information were read into SAS 9.4 for comparison. Convergence rate was analyzed first, then only the first 100 converged replications within each simulation condition were selected for further analysis. This chapter evaluates the performance of the three methods by comparing the convergence rate, accuracy of parameter estimates (bias and RMSD), and accuracy of DIF detection (power and Type I error rate).

4.1.1 Data Generation Validation

A single dataset was generated first to test the data generation process. The test dataset chose the largest number of groups (80 groups) with 40% DIF items (8 items) and equal group mean abilities. In order to evaluate the parameter recovery efficiently, a group size of 5,000 was

used for simulation. A total number of 20 dichotomous item responses were generated using SAS 9.4 with only one replication. The simulation procedure is described below in detail.

4.1.1.1 Procedure

Step 1: Generate two 8 × 20 datasets *dif8_a* and *dif8_b* with add-on DIF values using *proc iml*. The eight rows represent groups, while each group has only one DIF item. The 20 columns represent items, and item 1-8 were set as DIF items.

Based on the design, a small uniform DIF (0, 0.4) was added to item discrimination and item difficulty parameters of item 1 on group 1. A small nonuniform DIF (0.3, 0.4) was added to item parameters of item 2 on group 2. A medium uniform DIF (0, 0.6) and a medium nonuniform DIF (0.5, 0.6) were added to item 3 on group 3 and item 4 on group 4, respectively. The same pattern was repeated for item 5-8. Thus, the item discrimination parameters in dataset *dif8_a* are as follows:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & .3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & .5 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & .3 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .5 & 0 & \cdots & 0 \end{bmatrix}$$

Similarly, the item difficulty parameters in dataset *dif8_b* are as follows:

$$\begin{bmatrix} .4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & .4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & .6 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & .6 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & .4 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & .4 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .6 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .6 & 0 & \cdots & 0 \end{bmatrix}$$

Step 2: Create a marco to generate the data

1) Loop 1: do rep = 1 to number of replications

In this case, the number of replications is one. *Proc iml* was used to generate a 1×20 vector from a lognormal distribution with location parameter equals to 0 and scale parameter equals to 0.3. This vector was repeated eight times to create a 8×20 matrix *a0*. Read dataset *dif8_a* into a matrix *dif8_a*. The final item discrimination values in dataset *item1_a* is the sum of matrix *a0* and *dif8_a*.

For item difficulty parameter, *proc iml* was used to generate a 1×20 vector from a normal distribution with mean equals to 0 and variance equals to 0.25. Repeating this vector eight times and add the matrix dif8_b, then a dataset *item1_b* was obtained.

The item discrimination and item difficulty parameter was converted to the slope and intercept parameter in the random item effects model, in which the slope parameter equals to item discrimination, and the intercept parameter equals to the product of item discrimination and item difficulty.

2) Loop 2: do grep = 1 to number of replication of groups

In this case, the reapplication of groups was $\frac{80}{8} = 10$. This study combined dataset *item1_a* with *item1_b* to dataset *combine1*. The group index was generated to track the group number. The group index equaled to $group0 + (\&grep - 1) * 8$, in which $group0$ is the ordered number of rows.

3) Loop 3: do person = 1 to group size

This study created dataset *combine2* from *combine1* by generating individual ability parameters from a random normal distribution with mean equals to 0 and variance equals to 1. Dataset *combine3* was created, in which the probabilities of item responses were generated using Fox's random item effects model (Equation 27) with individual abilities, slope parameters, and

intercept parameters. In addition, a random number was generated from a uniform distribution. Item response was assigned as 1 if the probability was larger than or equal to the random number and assigned as 0 if the probability was smaller than the random number.

4.1.1.2 Results

In order to evaluate the accuracy of parameter estimation, Group 1, 2, 3, and 4 were selected to represent four unique pattern of DIF items. Item responses from each group were analyzed separately using CFA model with robust ML estimation and probit link in Mplus 8. The obtained estimates of slope and intercept parameter were converted back to item difficulty estimates and item discrimination estimates in order to compare with the true parameter values. The true and estimated item parameters are presented in Table 2. All correlations between true and estimated item parameters are larger than or equal to 0.989

Table 2. Parameter Estimates

			Group 1		Group 2		Group 3		Group 4	
item	a(DIF)	b(DIF)	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
1	0.812	0.398 (+0.4)	0.893	0.754	0.804	0.367	0.861	0.397	0.826	0.370
2	0.998 (+0.3)	0.870 (+0.4)	0.994	0.879	1.352	1.243	0.946	0.912	0.994	0.874
3	0.763	-0.233 (+0.6)	0.757	-0.199	0.814	-0.276	0.770	0.422	0.744	-0.264
4	0.790 (+0.5)	0.468 (+0.6)	0.780	0.481	0.793	0.446	0.811	0.480	1.251	1.052
5	1.132	-1.547	1.171	-1.505	1.161	-1.519	1.085	-1.682	1.127	-1.553
6	1.476	-0.444	1.458	-0.425	1.419	-0.473	1.494	-0.424	1.497	-0.453
7	1.145	0.052	1.137	0.069	1.129	0.054	1.107	0.100	1.133	0.073
8	0.445	-0.869	0.446	-0.872	0.454	-0.892	0.444	-0.852	0.416	-0.944
9	0.592	-0.236	0.590	-0.214	0.619	-0.287	0.616	-0.191	0.587	-0.247
10	1.203	-0.547	1.218	-0.507	1.132	-0.545	1.200	-0.509	1.189	-0.560
11	0.732	-0.218	0.744	-0.170	0.744	-0.245	0.715	-0.230	0.730	-0.221
12	0.854	-1.194	0.880	-1.157	0.881	-1.212	0.848	-1.141	0.905	-1.197
13	0.490	0.599	0.474	0.633	0.484	0.584	0.524	0.650	0.473	0.611
14	0.882	0.481	0.886	0.515	0.849	0.441	0.811	0.537	0.882	0.473
15	0.890	-1.056	0.886	-1.033	0.894	-1.101	0.965	-1.013	0.850	-1.068
16	0.763	-0.165	0.746	-0.130	0.743	-0.214	0.727	-0.195	0.747	-0.143
17	0.695	-0.380	0.681	-0.395	0.695	-0.452	0.755	-0.312	0.625	-0.427
18	1.192	-0.507	1.190	-0.483	1.204	-0.548	1.235	-0.491	1.207	-0.489
19	0.988	0.835	0.919	0.912	0.961	0.794	0.984	0.860	0.959	0.871
20	0.570	0.139	0.551	0.188	0.547	0.140	0.581	0.124	0.558	0.173
mean	0.871	-0.158	0.870	-0.133	0.884	-0.185	0.874	-0.128	0.885	-0.153
corr			0.994	0.999	0.994	0.999	0.989	0.998	0.996	0.999

4.2 Convergence

The initial study design is to have 100 replications for each simulation condition. However, a relatively high non-convergence rate was discovered soon after the data generation started. Non-convergence or slow convergence mostly happened on datasets with no variance or very small variance on the binary responses across individuals. To ensure enough data for valid analysis, the total number of replications was increased to 150. The convergence rate was calculated based on

the 150 replications. Then the first 100 converged replications within each simulation condition were selected for further analysis regarding the parameter recovery and DIF detection.

As discussed in Section 3, the convergence was determined in two steps: algorithms failed to converge within 50,000 iterations in Mplus, and algorithms converged in Mplus but with poor trace plots and autocorrelation plots. The final convergence rate is presented in Table 3 below.

Table 3. Convergence Rate

Proportion of DIF	Group Mean Abilities	NG	GS	Random Item Effects	FIXED Alignment	FREE Alignment
20%	Equal	24	100	100.00%	100.00%	100.00%
20%	Equal	24	500	100.00%	100.00%	100.00%
20%	Equal	40	100	100.00%	100.00%	100.00%
20%	Equal	40	500	100.00%	100.00%	100.00%
20%	Equal	80	100	100.00%	99.33%	96.67%
20%	Equal	80	500	95.33%	98.00%	96.00%
20%	Unequal	24	100	94.00%	95.33%	94.67%
20%	Unequal	24	500	92.00%	91.33%	90.00%
20%	Unequal	40	100	92.67%	94.67%	92.67%
20%	Unequal	40	500	94.67%	91.33%	90.00%
20%	Unequal	80	100	83.33%	82.00%	76.67%
20%	Unequal	80	500	91.33%	86.67%	84.00%
40%	Equal	24	100	100.00%	100.00%	100.00%
40%	Equal	24	500	100.00%	100.00%	100.00%
40%	Equal	40	100	100.00%	100.00%	100.00%
40%	Equal	40	500	100.00%	100.00%	100.00%
40%	Equal	80	100	99.33%	98.67%	92.67%
40%	Equal	80	500	90.00%	93.33%	90.67%
40%	Unequal	24	100	97.33%	95.33%	95.33%

40%	Unequal	24	500	93.33%	90.00%	89.33%
40%	Unequal	40	100	98.67%	98.00%	97.33%
40%	Unequal	40	500	97.33%	96.00%	94.00%
40%	Unequal	80	100	92.00%	88.67%	86.00%
40%	Unequal	80	500	90.67%	89.33%	88.67%

Among the 24 simulation conditions with 150 replications each, majority of conditions have a convergence rate higher than 90%. Lowest convergence rate (82%) was detected in conditions with 20% DIF items, unequal group mean abilities, 80 groups, and 100 individuals per group. According to Table 3, every algorithm converged when the group mean abilities were equal and the number of groups was not the largest (24 or 40 groups), despite the proportion of DIF items, group size, and estimation method. The reason is on the data generation level. Individual abilities in the equal group mean abilities conditions were generated from a normal distribution with a mean of 0 and a variance of 1. Individual abilities in the unequal group mean abilities conditions were generated from a normal distribution with a group mean of β_g and a group variance of 1, where β_g was generated from a normal distribution with a mean of 0 and a variance of 1 $\mathrm{N}(0, 1)$. In some unequal group mean abilities conditions with extremely high or extremely low group mean β_g, there is a higher chance of uniform responses or low variance responses due to the limited information provided by the binary responses. From this perspective, exclude non-converged data from further analysis is in fact excluding extreme group mean abilities from the study.

In equal group mean abilities conditions, the non-convergence rate increased as the total sample size increases. Simulated datasets started to show non-convergence when the number of groups increased to 80. When there were 80 groups with equal group mean abilities, larger non-convergence rates were found in conditions with large group size (GS=500). Similar pattern can

be detected among unequal group mean abilities conditions. The non-convergence rate increased as the total sample size increased. There was no clear discrepancy on convergence rates between conditions with 24 groups and conditions with 40 groups, but the convergence rate dropped below 92% as the number of groups increased to 80. When the number of groups was 80, increase the group size would generally increase the probability of convergence, except for the random item effect model in conditions with 40% DIF items and unequal group mean abilities. However, such consistent pattern could not be detected from conditions with smaller number of groups.

In general, the random item effect model tended to have a higher convergence rate than the alignment method, especially in conditions with large sample size. The only two conditions where the alignment method converged more is when the group mean abilities were unequal, the proportion of DIF items was 20%, the group size was 100, and the number of groups were smaller than 80. The convergence rates between the FIXED alignment method and the FREE alignment method were close, and the FIXED alignment method consistently converged more than the FREE alignment method. Those finding show that the FREE alignment method is most sensitive to non-ideal dataset, which makes it a more appropriate initial screening method for empirical data.

4.3 Accuracy of Parameter Estimation

Each simulated dataset was estimated using three methods: the random item effect model, the FIXED alignment method, and the FREE alignment method. The accuracy of parameter estimates was assessed by two indices: true bias and RMSD. This section compares the bias and RMSD across three estimation methods in varying simulation conditions.

4.3.1 Bias

Mixed ANOVA was used to evaluate the impact of four between subject factors (proportion of DIF items, group mean abilities, number of groups, and group size) and one within subject factor (estimation methods) on the true bias of parameters. Table 4 presents those factors and their levels.

Table 4. Simulation Parameters and Levels

Between Subject Factors	**Levels**
Proportion of DIF (dif)	20% (dif = 4)
	40% (dif = 8)
Group mean abilities (meanab)	Equal (meanab = 1)
	Unequal (meanab = 2)
Number of groups (ng)	24
	40
	80
Group size (gs)	100
	500
Within Subject Factors	**Levels**
Estimation methods (method)	Random item effects model (method = 1)
	FIXED alignment method (method = 2)
	FREE alignment methods (method = 3)

In general, the estimation of the slope parameter was more accurate than the estimation of the intercept parameter. All three estimation methods tended to overestimate the slope parameter in majority of conditions with biases ranged from -0.36 to 0.44. The overall mean and standard deviation of the biases of slope parameter across all conditions and replications are 0.07 and 0.10, respectively. The intercept parameter tended to be underestimated in general, but this pattern was not consistent across simulation factors. As compared to the biases of slope parameters, the biases of intercept parameters have a larger variance with mean equals to -0.04 and standard deviation equals to 0.26. The biases of intercepts range from -1.64 to 0.92. The large proportion of negative values makes the overall mean of biases less informative.

The main effects of five factors and two-way interactions are presented in Table 5 below, and only significant effects ($p < .05$) were reported. The assumption of sphericity was violated for all tested criteria, so the Greenhouse-Geisser correction was used. Significant main effect of proportion of DIF, group mean abilities, group size, and estimation methods on the estimation of slope parameters were found after controlling for other factors. Group size had a large significant impact on the estimation of slope parameters across all estimation methods, and this impact differed between estimation methods. Significant main effect of proportion of DIF, group mean abilities, number of groups, and group size were also detected on the estimation of intercept parameters, but none of them had a large effect size. All the significant between subject effects and within subject effects will be further discussed in the following sections.

Table 5. Overall Mixed ANOVA Results for Biases

		Bias of slope			Bias of intercept		
	DF	F	p	η_p^2	F	p	η_p^2
Proportion of DIF	(1,2394)	15.37	<.001	0.01	150.26	<.001	0.06
Group mean abilities	(1,2394)	6.05	0.01	< .01	11.45	<.001	< .01
Number of groups	(2,2394)				39.68	<.001	0.03
Group size	(1,2394)	2922.22	<.001	0.55	50.50	<.001	0.02
Method	(2,4788)	1332.29	<.001	0.36			
Proportion of DIF × Method	(2,4788)				30.97	<.001	0.01
Group mean abilities × Method	(2,4788)	4.45	0.01	< .01	6.76	<.01	< .01
Number of groups × Method	(4,4788)	9.99	<.001	0.01	20.02	<.001	0.02
Group size × Method	(2,4788)	617.26	<.001	0.20	16.46	<.001	0.01

4.3.1.1 Overall Comparison Between Three Methods

The biases of slope parameter show that all three estimation methods tended to overestimate the slope parameter. The overall mean bias of the slope parameter estimated by the random item effects

model, the FIXED alignment method, and the FREE alignment method were 0.02, 0.10, and 0.11, respectively. While other factors were controlled, the random item effect model was significantly more accurate than the two alignment methods on estimating the slope parameter, $F(1,2394) = 2238.08, p < .001, \eta_p^2 = .48$, and $F(1,2394) = 2520.10, p < .001, \eta_p^2 = .51$, respectively. In addition, the biases of slope parameter estimated by the random item effects model varied less across conditions and replications, indicates that the random item effects model was consistently more accurate than the alignment methods on estimating slope parameter regardless of other influential factors. The overall standard deviation of bias of slopes estimated by the three methods were: 0.03 for the random item effects model, 0.11 for the FIXED alignment method, and 0.11 for the FREE alignment method. A small significant difference on the biases of slope parameter was detected between the two alignment methods, but the effect size was too small to generalize this finding to empirical studies.

As compared to the estimation of slope parameter, the estimation of intercept parameter made by all three methods were generally less accurate with biases larger in absolute values, ranged from -1.64 to 0.92. Because a large proportion of the true biases were negative values, the overall mean of bias gives limited information about the performance of each estimation method. While other factors were controlled, no significant difference was detected in the biases of intercept across estimation methods (Table 5). However, the biases of intercept parameter made by the alignment methods had a noticeable larger variance than those made by the random item effects model. The standard deviation of the bias of intercept parameter made by the three estimation methods were: 0.03 for the random item effect model, 0.36 for the FIXED alignment method, and 0.27 for the FREE alignment method.

Overall, all three estimation methods were fairly accurate on estimating the slope and intercept parameters, while they all tended to overestimate the slope parameter. The random item effects model performed significantly different with the two alignment methods in the estimation of slope parameter after controlling for other factors. The interaction between estimation method and other simulation factors will be discussed in the following sections.

4.3.1.2 Proportion of DIF Items

Mixed ANOVA results show that the proportion of DIF items had a significant but limited impact on the accuracy of parameter estimates. Although significant main effect of the proportion of DIF was found on both the bias of slope parameter and the bias of intercept parameter (Table 5), this impact on the slope parameter was extremely small, and no significant interaction was detected between the proportion of DIF items and the estimation method. Given the small effect size, it can be concluded that the proportion of DIF items made no difference in estimating the slope parameters, regardless of the estimation methods.

A relatively larger impact of the proportion of DIF items was detected on the estimation of intercept parameter (Table 5). The absolute value of biases of intercept parameter tended to increase as the proportion of DIF items increased. This general trend was consistent across three estimation method, but the level of impact was different. Also, there was a tendency of underestimation in conditions with 40% DIF items (Figure 6).

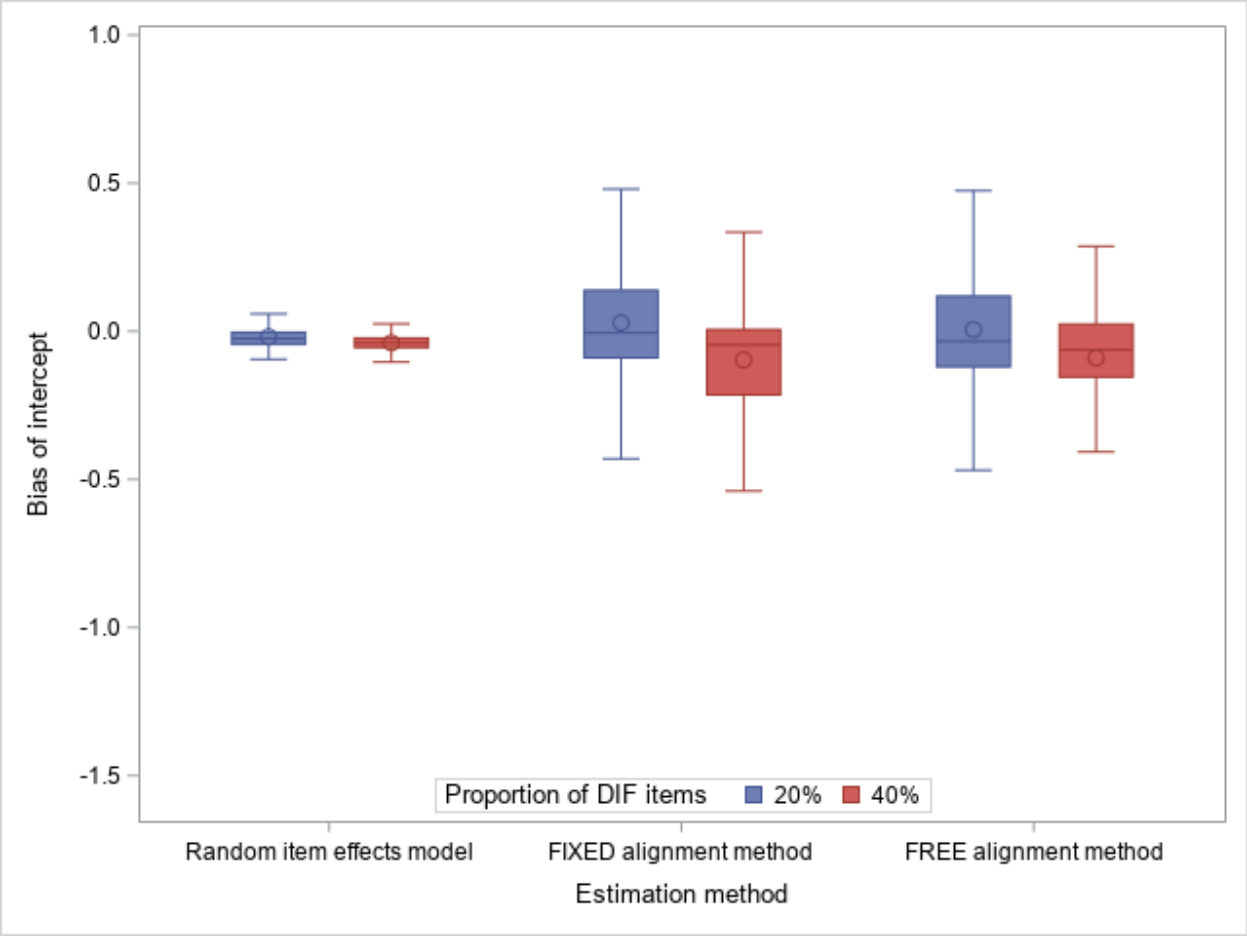

Figure 6. Boxplot of Biases of Intercept: Proportion of DIF Items × Estimation Method

A small significant interaction effect was detected between the estimation method and the proportion of DIF items, $F(2,4788) = 30.97, p < .001, \eta_p^2 = .01$. Although the effect size was small, a clear pattern was found regarding the impact of DIF proportion across estimation methods. The proportion of DIF items had a greater impact on the intercept parameter estimates made by the random item effects model than those made by the alignment methods (Table 6). When the random item effects model was used, the underestimation of intercept parameters increased significantly as the proportion of DIF items increased.

Table 6. The Effect of Proportion of DIF Items on Bias of Intercept by Estimation Method

Method	F	p	η_p^2
Random item effects model	276.62	< .001	0.10
FIXED alignment method	76.37	< .001	0.03
FREE alignment method	82.30	< .001	0.03

4.3.1.3 Group Mean Abilities

The type of group mean ability made no significant impact on the bias of slope parameter, regardless of the estimation method. A small significant difference was detected in the biases of intercept parameter between different types of group mean abilities (Table 5). The interaction between the group mean abilities and estimation method was small statistically, $F(2,4788) = 6.76, p < .01, \eta_p^2 < .01$, but the boxplot (Figure 7) below shows a clear pattern of this interaction effect. When the random item effects model was used for estimation, the accuracy of intercept parameter estimation was higher in conditions with unequal group mean abilities, $F(1,2394) = 169.25, p < .001, \eta_p^2 = .07$. When the alignment methods were used for estimation, the variance of biases of intercept increased dramatically in conditions with unequal group mean abilities (Figure 7), which indicates that the estimates of intercept are unreliable in those conditions.

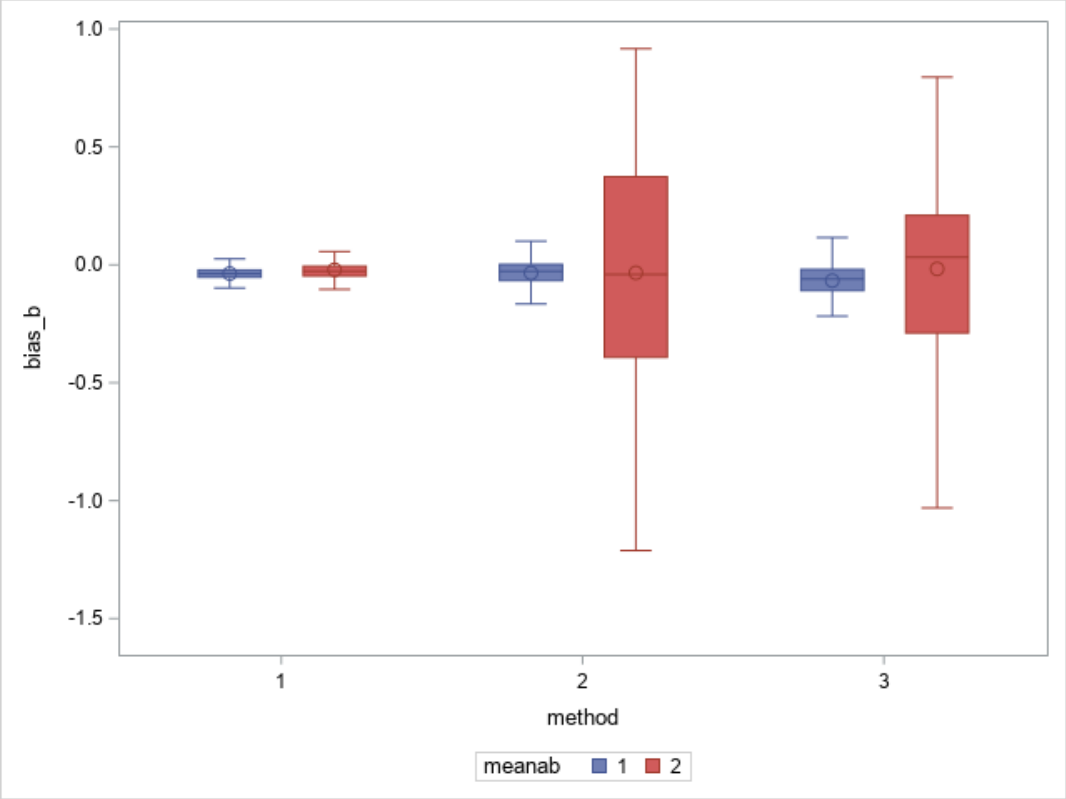

Figure 7. Boxplot of Biases of Intercept: Type of Group Mean Ability × Estimation Method

In conditions with equal group mean abilities, the random item effects model performed similar with the FIXED alignment method in terms of the mean biases of intercept parameter, and both methods were fairly accurate with mean bias of intercept around -0.03. However, the FREE alignment method was significantly less accurate than those two methods with a larger absolute mean bias and a larger variance of biases, $F(1,1195) = 199.79, p < .001, \eta_p^2 = .14$, $F(1,1195) = 184.23, p < .001, \eta_p^2 = .13$, respectively.

4.3.1.4 Number of Groups

Number of groups made no significant impact on the estimation of slope parameter (Table 5). The overall mean of biases of slope parameter was 0.07, and the standard deviation was 0.10.

In other words, the estimation of slope parameter was equally accurate across different numbers of groups.

In terms of the bias of intercept parameter, a small significant difference across different numbers of groups was detected (Table 5). It has been found that the estimation of intercept parameter was least stable in conditions with the smallest number of groups, where the standard deviation equaled to 0.32. Although some significant differences were found from the multiple comparison results (Table 7), the overall interaction effect between the number of groups was very small ($\eta_p^2 = .02$). In general, the random item effects model tended to be more accurate than the alignment methods on estimating the intercept with smaller absolute value of biases and smaller variance of biases, regardless of number of groups.

Table 7. Multiple Comparison Result: Number of Groups × Estimation Method

	24 vs. 40 groups	24 vs. 80 groups	40 vs. 80 groups
Random item effects model	$p = .04$	$p < .001$	$p < .001$
FIXED alignment method	$p < .001$		$p < .001$
FREE alignment method	$p < .001$	$p < .001$	$p < .001$

4.3.1.5 Group Size

Group size made a significant impact on the estimates of slope parameter, and this impact differed significantly across estimation methods after controlling for other factors (Table 5, Figure 8). There was a consistent pattern across all three methods that the estimation accuracy of slope parameter increased significantly as the group size increased.

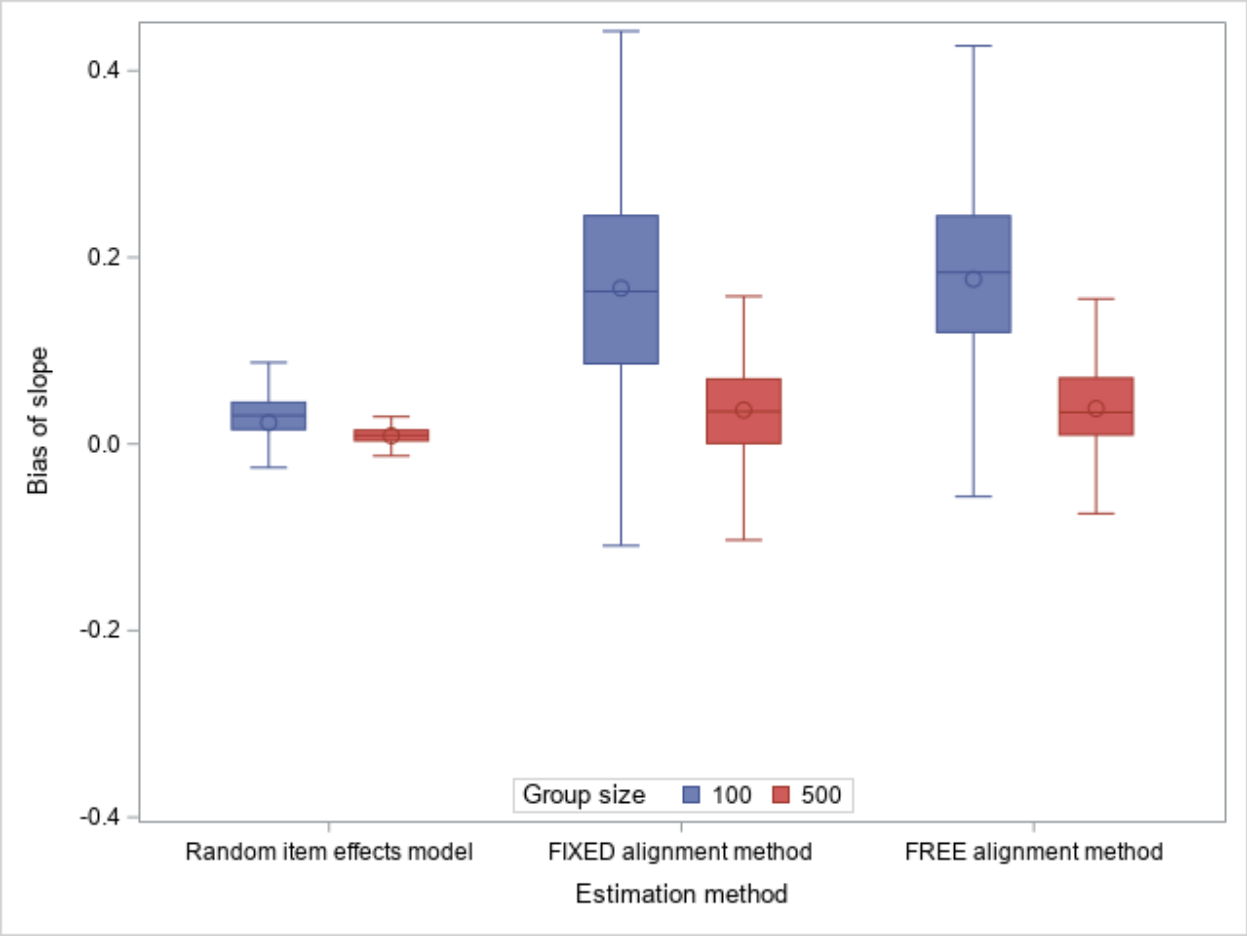

Figure 8. Boxplot of Biases of Slope: Group Size × Estimation Method

It has been found that group size had a significantly larger impact on the estimates of slope parameter made by the two alignment methods than those by the random item effects model (Table 8). In conditions with 100 individuals per group, the random item effects model was significantly more accurate on estimating slope parameter as compared to the two alignment methods, $F(1,1195) = 1906.39,\ p < .001,\ \eta_p^2 = .61$,$F(1,1195) = 2158.99,\ p < .001,\ \eta_p^2 = .64$, respectively. As the group size increased, the biases of slope made by the alignment methods increased more dramatically than those made by the random item effects model. As a result, the discrepancy in biases of slope parameter between the random item effects model and the alignment methods decreased. In conditions with 500 individuals per group, the biases of slope made by the

random item effects model was still more accurate than those made by the alignment methods with smaller mean and variance. But the discrepancy between estimation methods decreased with smaller significant differences between the random item effects model and the two alignment methods, $F(1,1195) = 371.00,\ p < .001,\ \eta_p^2 = .24,\ F(1,1195) = 475.63,\ p < .001,\ \eta_p^2 = .28$, respectively. In terms of the two alignment methods, they performed similarly across conditions with only a small significant difference ($\eta_p^2 < .01$) between them.

Table 8. The Effect of Group Size on Bias of Slope by Estimation Method

Method	F	p	η_p^2
Random item effects model	125.23	< .001	0.05
FIXED alignment method	1390.48	< .001	0.37
FREE alignment method	1731.93	< .001	0.42

The group size made a small significant impact on the estimation of intercept parameter, and this impact differed significantly across estimation methods (Table 9). Due to the small effect size in Table 9, it can be concluded that the group size made no difference on the bias of the intercept parameter, regardless of the estimation method.

Table 9. The Effect of Group Size on the Bias of Intercept

Method	F	p	η_p^2
Random item effects model	1.59	-	-
FIXED alignment method	19.53	< .001	0.01
FREE alignment method	50.76	< .001	0.02

4.3.2 RMSD

This section discusses the accuracy of parameter recovery by evaluating the RMSD of expected score, slope parameter, and intercept parameter. It has been found that the overall parameter recovery for slope, intercept, and expected score were fairly accurate across all conditions with mean of RMSDs equals to 0.21, 0.34, and 0.44, respectively. In consistent with the analysis of bias, the estimates of slope parameter were generally more accurate than the estimates of intercept parameter with smaller mean and standard deviation of RMSDs.

Mixed ANOVA results of main effects and interaction effects are presented in Table 10 below. The Greenhouse-Geisser correction was used when the assumption of sphericity was violated. As compared to the analysis of bias, the analysis of RMSD is more likely to capture significant effects because overestimation and underestimation would not counterbalance each other across items and individuals. In addition to that, the RMSD of expected score can assess the overall model recovery and provide information that is comparable across studies. Table 10 shows that the type of group mean abilities and estimation method made the largest impact on the overall parameter recovery. Besides those two factors, the group size also had a large significant impact on the estimation of slope parameter. All significant between subject effects and within subject effects will be further discussed in this section.

Table 10. Overall Mixed ANOVA Results for RMSDs

		RMSD of expected score			RMSD of slope			RMSD of intercept		
	DF	F	p	η_p^2	F	p	η_p^2	F	p	η_p^2
Proportion of DIF	(1,2394)				45.07	<.001	0.02			
Group mean abilities	(1,2394)	1491.60	<.001	0.38	91.25	<.001	0.04	3232.74	<.001	0.57
Number of groups	(2,2394)	14.75	<.001	0.01	8.68	<.001	0.01	77.53	<.001	0.06
Group size	(1,2394)				6579.34	<.001	0.73	42.69	<.001	0.02
Method	(2,4788)	2503.34	<.001	0.51	1322.18	<.001	0.36	763.34	<.001	0.24

Proportion of DIF × Method	(2,4788)				57.64	<.001	0.02	17.19	<.001	0.01
Group mean abilities × Method	(2,4788)	1931.16	<.001	0.45	10.75	<.001	< .01	925.01	<.001	0.28
Number of groups × Method	(4,4788)	21.30	<.001	0.02	11.54	<.001	0.01	28.29	<.001	0.02
Group size × Method	(2,4788)				363.14	<.001	0.13	16.19	<.001	0.01

There are some inconsistencies between the conclusions draw from RMSDs and the conclusions draw from biases in Section 4.3.1, especially on the estimation of intercept parameters. The main reason is that biases have negative values. After averaging across the 20 items, the bias of each parameter would move towards 0, and it was more likely to be affected by extreme values. The results about slope parameter are better aligned with each other because the slope parameter tended to be overestimated with a small proportion of negative biases. The intercept parameter tended to be underestimated in general with a large proportion of positive biases, and the variance of biases was much large as compared to the slope parameter. As a result, the analysis intercept estimates leaded to very different conclusions.

4.3.2.1 Overall Comparison Between Three Estimation Methods

The random item effects model performed significantly different with the alignment methods according to the RMSDs of expected score, slope, and intercept parameter. The Mixed ANOVA analysis on the RMSD of expected score demonstrated that the two alignment methods were significantly more accurate than the random item effect model in terms of the overall parameter recovery after controlling for other factors, $F\ (1, 2394) = 2497.78, p < .001, \eta_p^2 = .51$, and $F\ (1, 2394) = 2540.16, p < .001, \eta_p^2 = .51$, respectively. The mean RMSD of expected score estimated by the random item effect model, FIXED alignment method, and FREE alignment method across all simulation conditions were 0.62, 0.36, and 0.36, respectively. The performance of the two alignment methods were similar with only a small significant difference ($\eta_p^2 < .01$).

The random item effects model was more accurate than the two alignment methods on estimating the slope parameter. The mean RMSDs of slope parameter made by the random item effects model have a smaller mean and a smaller variance as compared to those made by the two alignment methods (Figure 9). It needs to be noted that 84 outliers (RMSD > 2) were excluded from Figure 9 to better presents the mean differences between three estimation methods, but those outliers were not excluded from any statistical analysis. While other factors were controlled, the random item effects model was significantly more accurate than the FIXED and FREE alignment method, $F\ (1, 2394) = 1380.82,\ p < .001,\ \eta_p^2 = .37$, and $F\ (1, 2394) = 1559.11,\ p < .001,\ \eta_p^2 = .39$, respectively. The two alignment methods performed similar with only a small significant difference ($\eta_p^2 < .01$) between them.

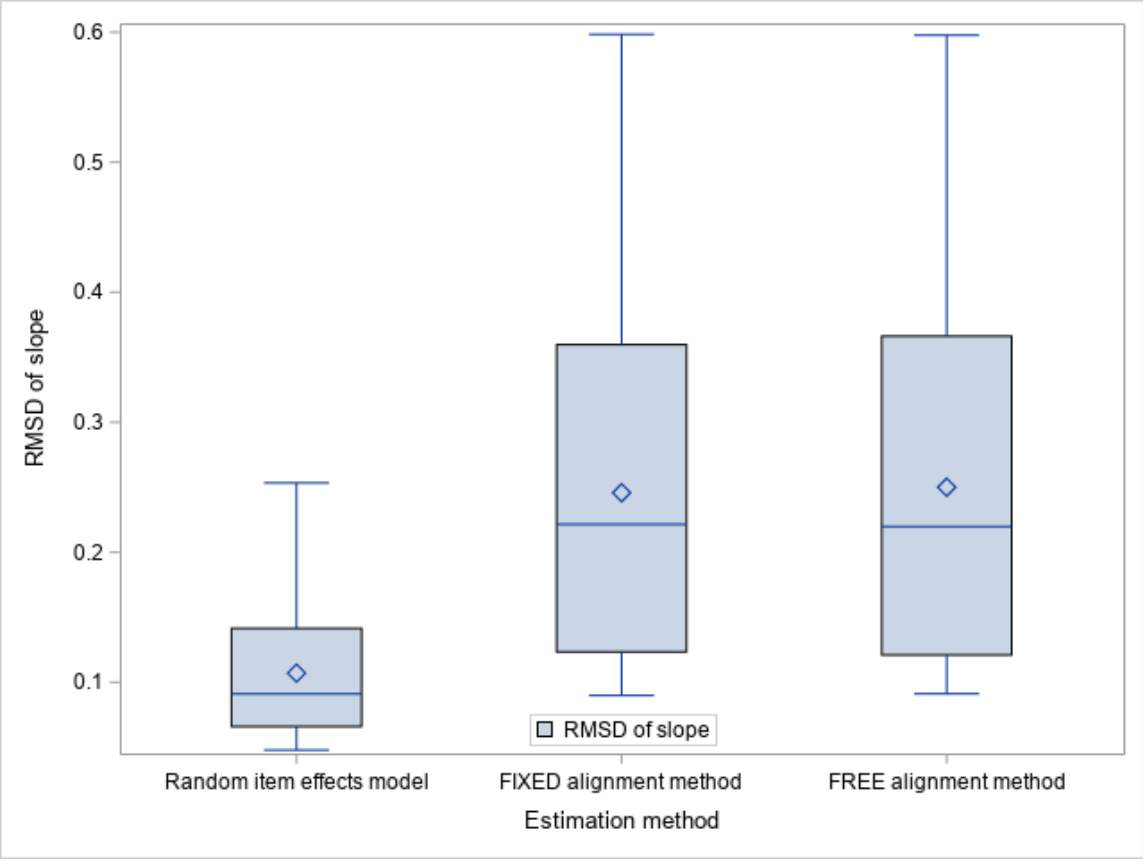

Figure 9. Boxplot of RMSD of Slope by Estimation Method

The random item effects model performed significantly different from the two alignment methods with regard of intercept parameter estimation, $F\ (1, 2394) = 797.15, p < .001, \eta_p^2 = .25$, and $F\ (1, 2394) = 1929.96, p < .001, \eta_p^2 = .45$, respectively. Although the RMSDs of intercept made by the random item effects model tended to be larger, they were more stable with a smaller standard deviation as compared to those made by the two alignment methods (Figure 10). The RMSDs of intercept made by the alignment methods had smaller means, but the estimation can sometimes be highly inaccurate with extremely large RMSDs. Between the two alignment methods, the FIXED alignment method seemed to be less accurate with a larger variance of RMSDs of intercept. But in fact, only a small statistical difference ($\eta_p^2 = .03$) was detected between those two methods.

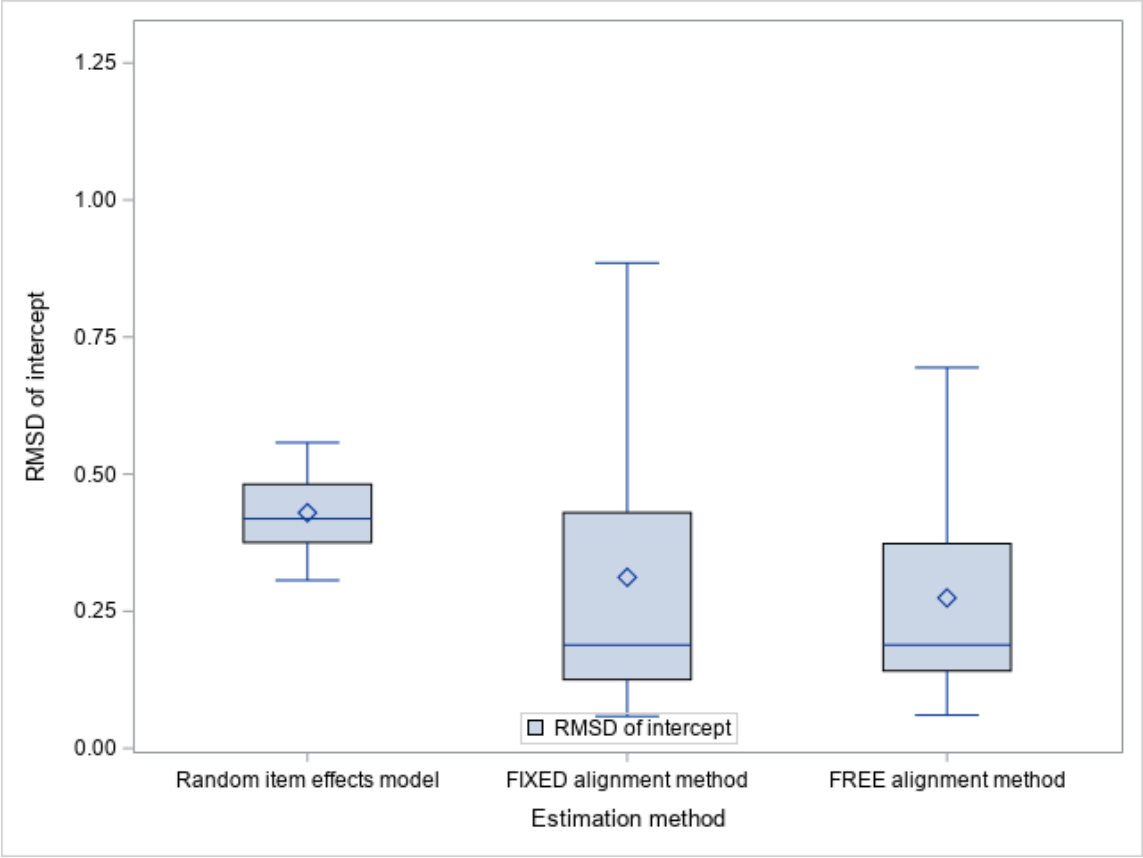

Figure 10. Boxplot of RMSD of Intercept by Estimation Method

4.3.2.2 Proportion of DIF Items

The proportion of DIF items made no significant impact on the RMSDs of expected score or on the RMSDs of intercept. But it had a small significant impact on the RMSDs of slope when random item effects model was used for parameter estimation (Table 10). In conditions where the random item effects model was used for estimation, the estimation accuracy increased as the proportion of DIF items increased, $F\ (1, 2394) = 67.36, p < .001, \eta_p^2 = .03$. In conditions where the alignment methods were used for estimation, no significant difference was detected on RMSD of slopes between the 20% DIF items conditions and the 40% DIF items conditions.

4.3.2.3 Group Mean Abilities

The type of group mean abilities had a significant impact on the estimation of expected score made by all three methods (Table 10). It had a greater impact on the estimation made by the random item effects model than those made by the alignment methods (Table 11). All three estimation were fairly accurate in conditions with equal group mean abilities with the mean RMSD of expected score equals to 0.40, 0.37, 0.37, respectively. The FIXED and FREE alignment methods were significantly more accurate than the random item effects model on estimating expected scores in those conditions, $F\ (1,1195) = 257.79, p < .001, \eta_p^2 = .18,$ and $F\ (1, 1195) = 262.15, p < .001, \eta_p^2 = .18$, respectively. The estimation accuracy of the expected score made by the two alignment methods decreased significantly in conditions with unequal group mean abilities. Similarly, the random item effects model became highly inaccurate in conditions with unequal group mean abilities. In those conditions, the mean and variance of RMSDs of expected score made by the random item effects model were 0.83 and 0.36, respectively. The inaccurate estimates of expected score were largely caused by the inaccurate estimates of slope,

which will be discussed in this section. In terms of the performance between the two alignment methods, they again showed similar pattern in both conditions with only a small significant difference ($\eta_p^2 < .01$).

Table 11. The Effect of Group Mean Abilities on the RMSD of Expected Score

Method	F	p	η_p^2
Random item effects model	1802.47	< .001	.43
FIXED alignment method	563.40	< .001	.19
FREE alignment method	592.12	< .001	.20

The type of group mean abilities had a small significant impact on the estimates of slope parameter after controlling for other factors, and this effect differed significantly across estimation methods (Table 10). For all three estimation methods, the estimates of slope were more accurate in conditions with equal group mean abilities. It has been found that the type of group mean abilities had a slightly greater impact on the alignment methods than on the random item effects model, but the effect sizes were too small to generalize this result. While other factors were controlled, the two alignment methods performed similarly on estimating the slope parameters with only a marginal significant difference, $F\ (1, 2394) = 3.31, p = .07, \eta_p^2 < .01$. The RMSDs of slope estimated by the random item effects model were significantly smaller with less variance than those estimated by the alignment methods in both equal group mean ability conditions and unequal group mean ability conditions, $F\ (2, 2390) = 4188.82, p < .001, \eta_p^2 = .78$, $F\ (2, 2390) = 310.11, p < .001, \eta_p^2 = .21$. However, the distribution of raw RMSDs showed that the RMSDs of slope tended to have extreme values in conditions with unequal group mean abilities (Figure 11). It has been confirmed that the extreme RMSDs of slope all came from conditions with unequal group mean abilities, 20% DIF items, 80 groups, and 100 individuals per

group. Further analysis found that the main reason for extreme large RMSDs of slope is the extreme large true *a* value on individual items. It indicates that the estimates of slope parameter made by the random item effects model could be highly unreliable in the above conditions.

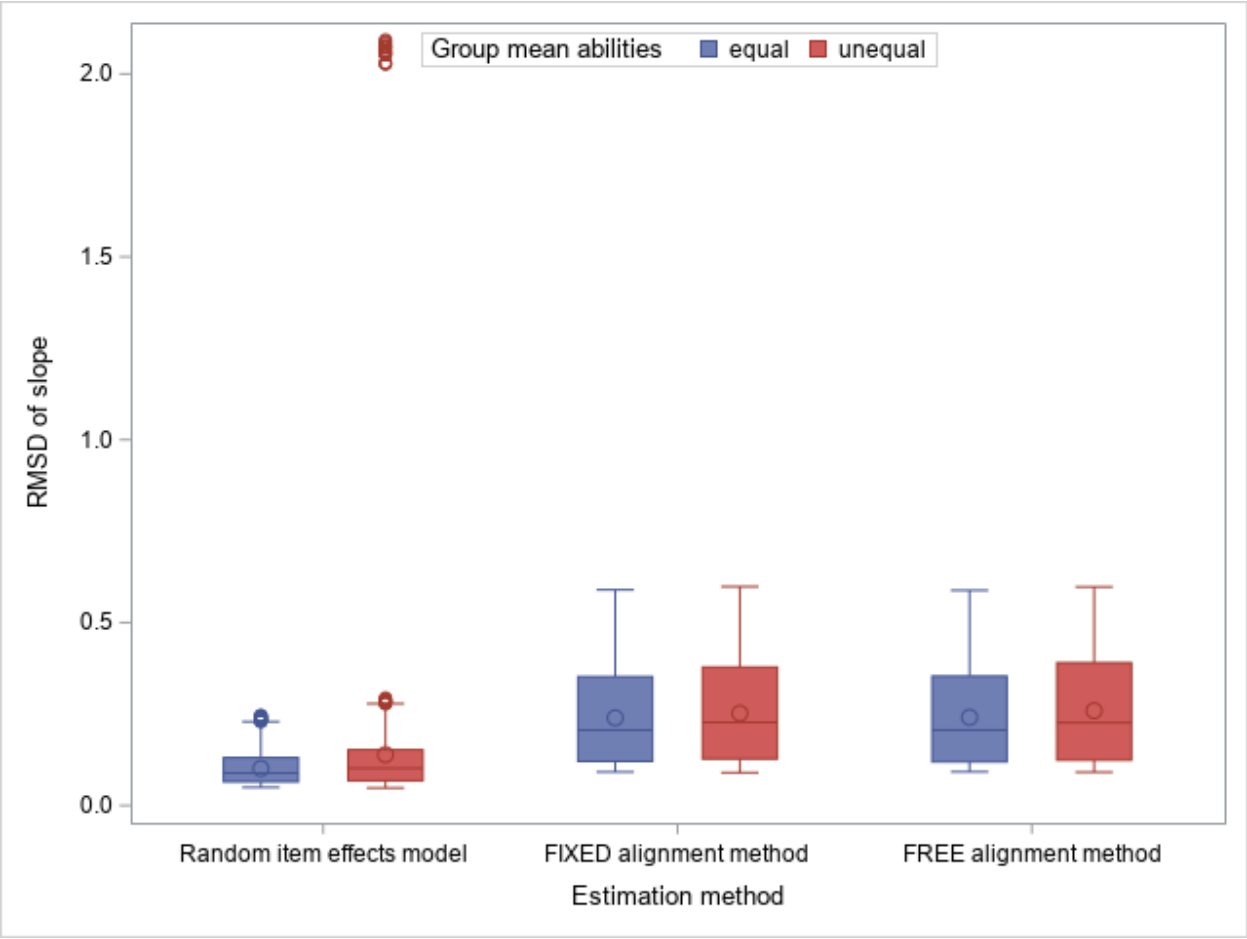

Figure 11. Boxplot of RMSD of Slope by Estimation Method × Group Mean Abilities

The significant main effect of group mean abilities mainly came from its impact on the estimation of intercept parameter (Table 10). The estimates of intercept made by the random item effects were generally not affected the type of group mean abilities, $\eta_p^2 < .01$. However, the type of group mean abilities had a large significant impact on the estimates of intercept made by the FIXED and FREE alignment method, $F\,(1, 2394) = 2174.93, p < .001, \eta_p^2 = .48$, and $F\,(1, 2394) = 1551.88, p < .001, \eta_p^2 = .39$, respectively. In conditions with equal group mean

abilities, the RMSDs of intercept made by the alignment were significantly smaller than those made by the random item effects model, $F\,(2, 2390) = 14738.10, p < .001, \eta_p^2 = .92$. In conditions with unequal group mean abilities, the RMSDs made by the random item effects model remained the same, but the RMSDs made by the two alignment methods increased significantly (Figure 12).

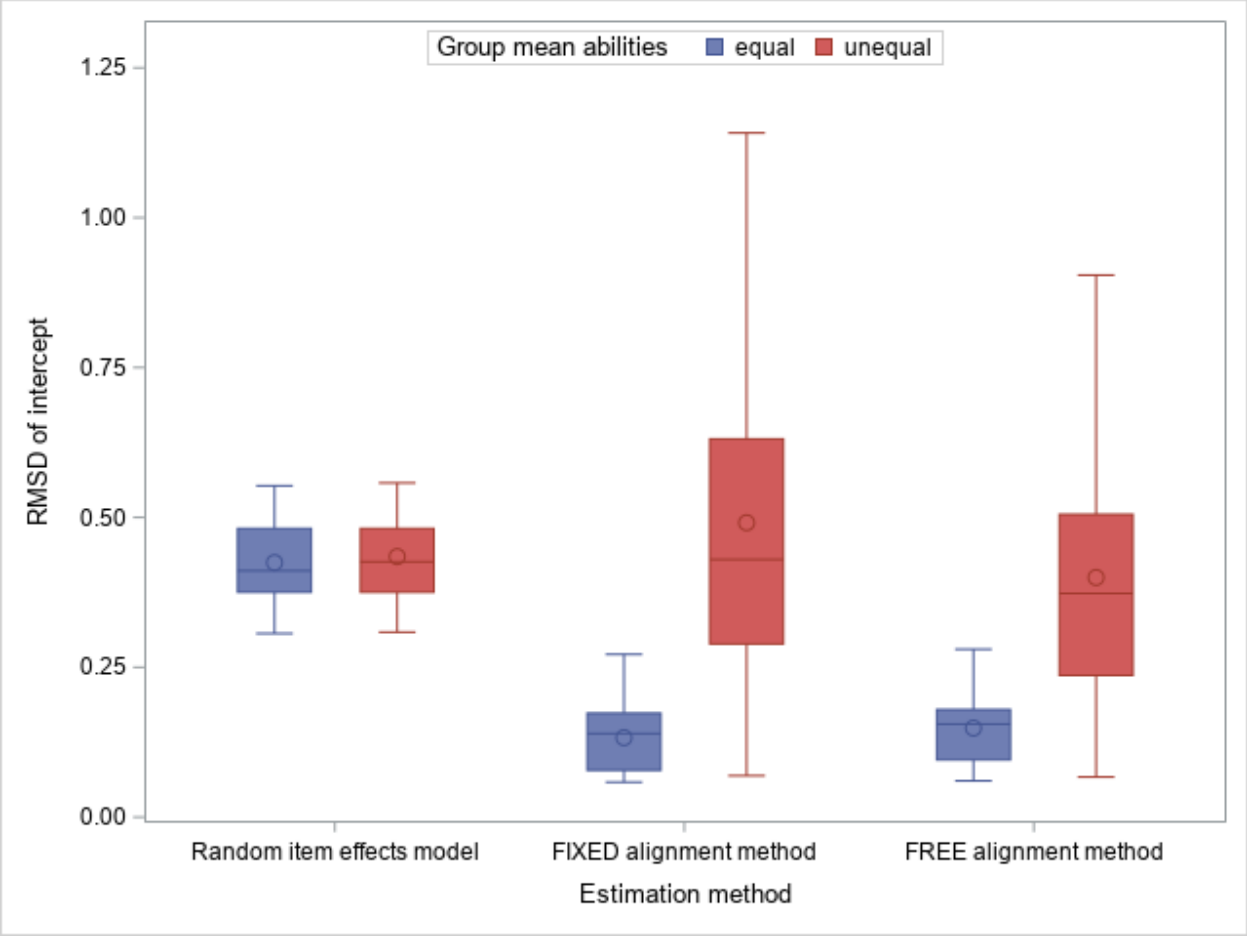

Figure 12. Boxplot of RMSD of Intercept by Estimation Method × Group Mean Abilities

In addition, a significant difference in the RMSDs of slope was detected between the two alignment methods, $F\,(1, 2394) = 132.45, p < .001, \eta_p^2 = .05$. In conditions with equal group mean abilities, the estimates of intercept made by the FIXED alignment method were significantly more accurate than those made by the FREE alignment method, $F\,(1, 1195) = 137.98, p <$

$.001, \eta_p^2 = .10$. But in conditions with unequal group mean abilities, despite that neither alignment methods was accurate, the FREE alignment method was more accurate than the FIXED one in terms of intercept estimates, $F\ (1, 1195) = 101.09, p < .001, \eta_p^2 = .08$.

4.3.2.4 Number of Groups

The number of groups made a small significant impact on the overall parameter recovery while controlling for other factors, and this impact differed across estimation methods (Table 10). When alignment methods were used for estimation, the RMSDs of expected score showed very similar pattern across different numbers of groups with only a small significant difference among them, η_p^2 equaled to .02, .01, and .01, respectively. In terms of the random item effects model, it was slightly more accurate in condition with 40 groups than in other conditions ($\eta_p^2 = .01$). The RMSDs of expected score had the smallest mean and variance in conditions with 40 groups as compared to in conditions with 24 and 80 groups.

The number of groups had an even smaller significant impact on the estimates of slope parameter after controlling for other factors (Table 10). This effect will not be discussed in detail due to the small effect size and unclear pattern. However, it needs to be noted that the RMSDs of slope made by the random item effects model in conditions with 80 groups had a larger number of extreme values than in conditions with smaller number of groups. It is mainly because the estimates of slopes were more likely to be inaccurate in conditions with large number of groups, small group size, and unequal group mean abilities. In conditions with 24, 40, and 80 groups, the estimates of slope made by the random item effects model were consistently more accurate than those made by the alignment methods, $F\ (2, 1592) = 2269.78, p < .001, \eta_p^2 = .74$, $F\ (2, 1592) = 2468.16$, $p < .001, \eta_p^2 = .76$, and $F\ (2, 1592) = 119.64, p < .001, \eta_p^2 = .13$, respectively.

The number of groups had no impact on the estimation of intercept made by the random item effects model, but it made a small significant impact on the estimation made by the alignment methods (Table 10). Despite the small effect sizes, the RMSDs of intercept made by the alignment methods decreased as the number of groups increased (Table 12).

Table 12. The Effect of Number of Groups on the RMSD of Intercept

Method	F	p	η_p^2
Random item effects model	0.05		
FIXED alignment method	66.33	< .001	.05
FREE alignment method	27.02	< .001	.03

Although the RMSDs of intercept made by the alignment methods had a smaller mean than those made by the random item effects model across conditions with different numbers of groups, the estimation was unreliable with a large variance of RMSDs in conditions with smaller number of groups (Figure 13). Between the two alignment methods, the increased number of groups had a slightly larger impact ($\eta_p^2 < .01$) on the estimates of intercept made by the FREE alignment method than those made by the FIXED alignment method.

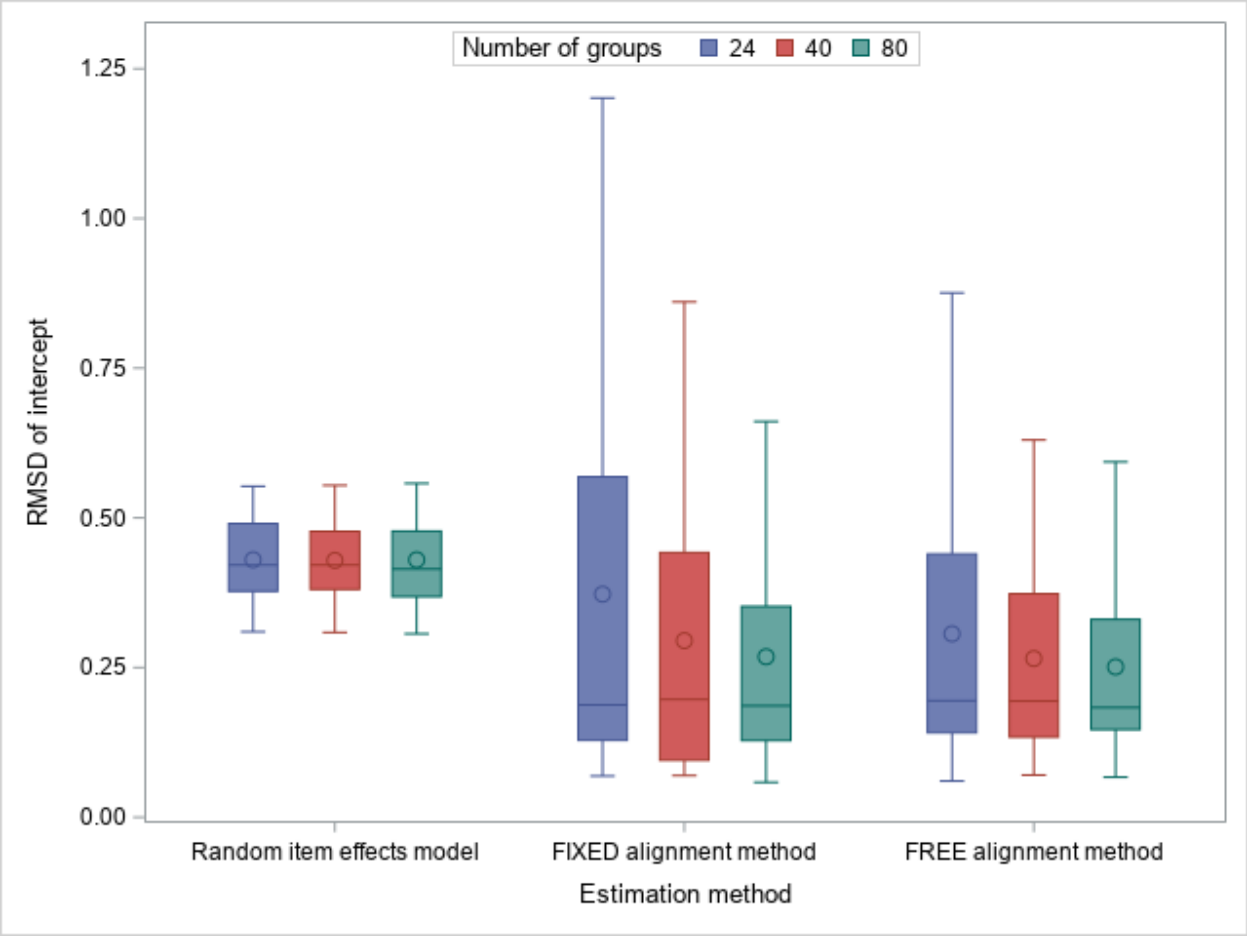

Figure 13. Boxplot of RMSD of Intercept by Estimation Method × Number of Groups

4.3.2.5 Group Size

The group size did not affect the overall recovery after controlling for other factors. It had a considerably large significant impact on the estimation of slopes, and a small significant impact on the estimation of intercepts (Table 10).

While other factors were controlled, the estimation accuracy of slope parameter increased significantly as the group size increased, regardless of estimation method (Table 10). The effect of group size had a significantly larger impact on the estimates of slope made by the alignment methods than those made by the random item effects model (Table 13).

Table 13. The Effect of Group Size on the RMSD of Slope

Method	F	p	η_p^2
Random item effects model	273.87	< .001	.10
FIXED alignment method	10561.2	< .001	.82
FREE alignment method	10999.5	< .001	.82

In conditions with 100 individuals per group, the random item effects model was significantly more accurate than the two alignment methods on estimating slope parameters, $F\ (1, 1195) = 828.60, p < .001, \eta_p^2 = .41,$ $F\ (1, 1195) = 965.82, p < .001, \eta_p^2 = .45,$ respectively. The RMSDs of slope made by the alignment methods had a larger mean and variance as compare to those made by the random item effects model (Figure 14). Again, extreme large RMSDs of slope were excluded from Figure 14, but they were included in all the analyses. As the group size increased to 500, the RMSDs of slope made by all three estimation methods became fairly accurate with mean equaled to 0.07, 0.13, and 0.13, respectively. The estimates of slope made by the random item effects model were still more accurate than those made by the two alignment methods in such conditions, $F\ (1, 1195) = 6745.26, p < .001, \eta_p^2 = .85,$ $F\ (1, 1195) = 6058.61, p < .001, \eta_p^2 = .84,$ respectively. The two alignment methods performed similarly on estimating slope parameters in conditions with both group sizes while only small significant differences were detected between them ($\eta_p^2 < .01$).

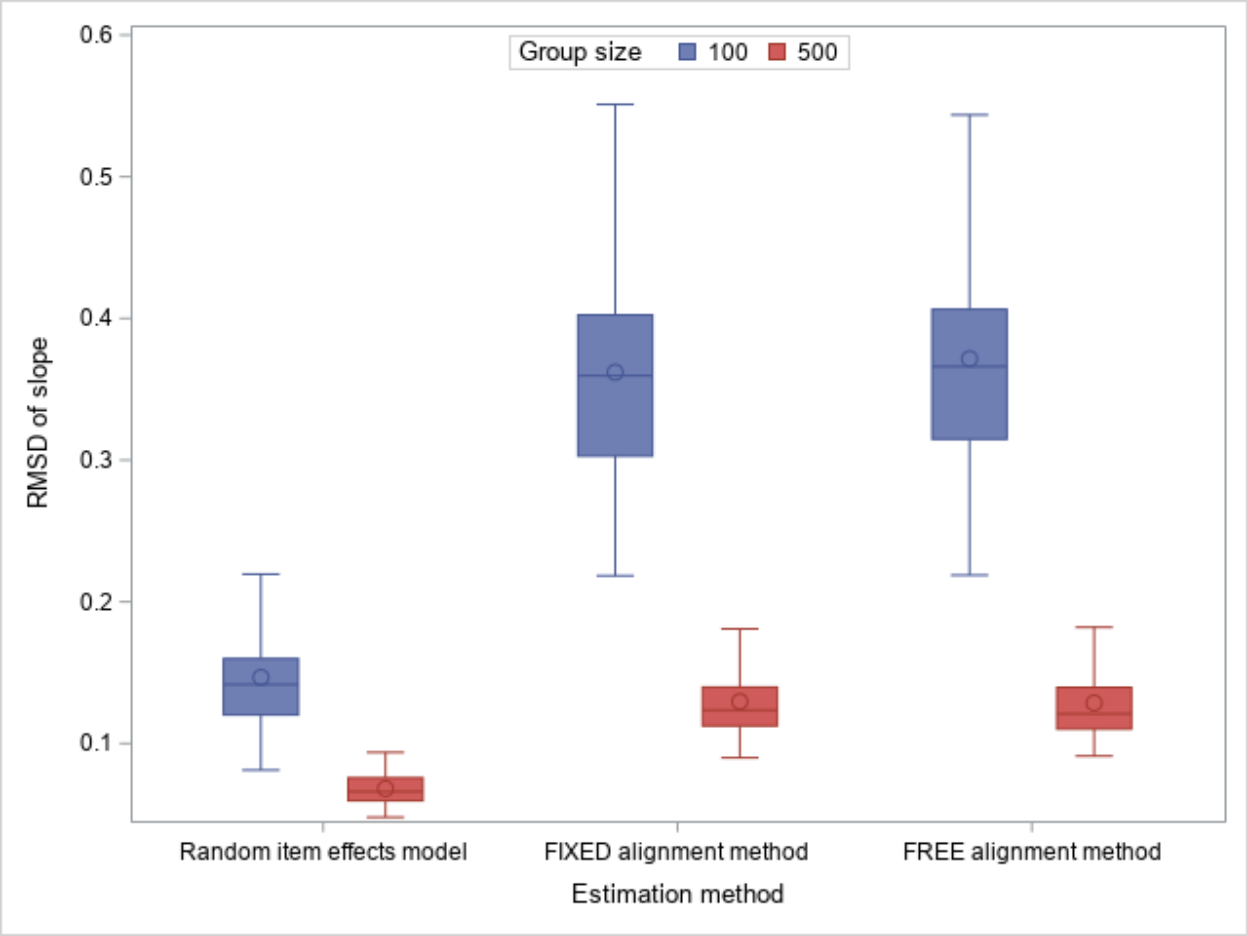

Figure 14. Boxplot of RMSD of Slope by Estimation Method × Group Size

The group size made a small significant impact on the estimates of intercept parameters after controlling for other factors, and this effect differed across estimation methods (Table 10). The main effect and interaction effect will not be discussed in detail due to the small effect size, except one important finding with regard of the two alignment methods. They performed similarly in conditions with 100 individuals per group with only a small significant difference ($\eta_p^2 < .01$). But in conditions with 500 individuals per group, the FIXED alignment method became significantly less accurate than the FREE alignment method on estimating intercept parameters, $F\,(1, 1195) = 69.87, p < .001, \eta_p^2 = .06$.

4.3.3 Summary

In general, all three estimation methods were accurate on the overall parameter recovery. When the proportion of DIF items, type of group mean abilities, number of groups, and group size were controlled, the two alignment methods were more accurate than the random item effect model, according to the RMSDs of expect score. On the individual item parameter level, the random item effects model was more accurate on estimating slope parameters, while the two alignment methods were more accurate on estimating intercept parameters. The random item effect model tended to overestimate slope parameters and underestimate intercept parameters, and the two alignment methods tended to overestimate both parameters. Small discrepancies between the two alignment methods can sometimes be detected, but the overall performance between the two methods were similar.

Number of groups, which is the primary interest of this study, had a limited impact on the accuracy of parameter estimation for all three methods. The random item effects model tended to be more accurate on the overall parameter recovery in conditions with 40 groups as compared to conditions with 24 and 80 groups. This finding is consistent with the general sample size requirement for the multilevel modeling. The analysis of number of groups demonstrated that: 1) the alignment methods performs equally well across different numbers of groups; and 2) the random item effects model will remain accurate as long as the level-2 sample size is large enough for a multilevel model.

The proportion of DIF items only made a small impact on the estimation of slope parameter, and it is difficult to generalize this finding to empirical studies due to the small effect size. The analysis of biases and the analysis of RMSDs of intercept leaded to different conclusion of the impact of DIF item proportion on intercept estimates. This is mainly because the positive

biases and negative biases of intercept on each individual item counterbalanced each other, and the extreme values of bias shifted the overall bias of intercept for each simulation condition. Asparouhov and Muthén (2014) claimed that large proportion of DIF items with small sample size tend to result in biased parameter estimates, but this study showed that the biased estimates are more likely a result of small group size rather than the large proportion of DIF items.

The type of group mean abilities and group size each both made large significant impact on the parameter estimation. In general, parameter estimates were more accurate in conditions with equal group mean abilities and larger group size after controlling for other factors. The performance of the three methods in conditions with equal group mean abilities is consistent with the conclusions from previous studies (De Boeck, 2008; Finch, 2016). This study confirmed that 500 individuals per group, as been demonstrated in previous studies (De Boeck, 2008; Finch, 2016; Stark, 2006), was enough for accurate parameter estimates in equal group mean abilities conditions, but 100 individuals per group would lead to biased parameter estimates. Asparouhov and Muthén (2014) found that 100 individual per group is sufficient for accurate parameter estimates when alignment method was used, but this study found that unequal group mean abilities and small group size together, would lead to inaccurate estimates of intercept when alignment methods were used for estimation. Such conditions had a greater impact on the random item effects model. Unequal group mean abilities and small group size would result in highly inaccurate estimates of slope parameter and inaccurate estimates of intercept parameter when the random item effects model was used. When the type of group mean abilities was controlled with other factors together, the estimation accuracy of slope parameter increased as the group size increased, regardless of the estimation method. In addition, the two alignment methods showed the largest difference between conditions with different group sizes as compared to other simulation factors.

The increased group size improved the estimation accuracy made by the FREE alignment method but made no impact on the estimation made by the FIXED alignment method.

4.4 Accuracy on DIF Detection

One of the major advantages of the alignment method is that it can flag items with DIF directly in the Mplus output. Section 4.2 and 4.3 have demonstrated that the two alignment methods performed very similarly on estimating slope and intercept parameters. This section will further explore the performance of FIXED and FREE alignment method regarding the accuracy of DIF detection. The power and Type I error were calculated by comparing the flagged DIF with the

As mentioned in the simulation design section, there are four combinations of DIF size and DIF type: (1) uniform DIF with small DIF size ($a + 0, b + 0.4$); (2) non-uniform DIF with small DIF size ($a + 0.3, b + 0.4$); (3) uniform DIF with medium DIF size ($a + 0, b + 0.6$); and (4) non-uniform DIF with medium DIF size ($a + 0.5, b + 0.6$). In conditions with 20% DIF items, Item 1 to Item 4 were set as uniform DIF with small DIF size, non-uniform DIF with small DIF size, uniform DIF with medium DIF size, and non-uniform DIF with medium DIF size correspondingly. In conditions with 40% DIF items, Item 1 to Item 4 were set in the same way as in 20% DIF items conditions, then Item 5 to Item 8 were set to repeat this pattern.

4.4.1 Power

The heat map below (Table 14) presents the power of all parameters with build-in DIF in conditions with 20% DIF items. Due to the limited space, slope parameter and intercept parameter were represented by *a* and *b*, respectively. Those symbols shall not be confused with the item discrimination parameter and item difficulty parameter in the IRT model. The following patterns can be found from this heat map.

1) When non-uniform DIF existed, the alignment methods were more accurate on detecting DIF in intercept parameters than on slope parameters. The DIF detection on slope parameter was highly inaccurate (power < 0.5) in almost all conditions. The power on slope parameters increased as the DIF size increased, but the overall power on slope parameters was still too small to be considered as reliable in empirical studies.
2) While other simulation factors were controlled, the DIF detection on both parameters was more accurate in conditions with 500 individuals per group than in conditions with 100 individuals per group. In conditions with small group size (GS = 100), the DIF detection on intercept parameters was highly inaccurate (power < 0.5) unless the number of groups was large enough (NG = 80). In conditions with large group size (GS = 100), a minimum power at 0.6 can be found from all intercept parameters regardless of other simulation factors.
3) In general, the DIF detection was most accurate in conditions with equal group mean abilities and large group size. Largest powers can be found in the condition with equal group mean abilities, 40 groups, and 500 individuals per group. In this condition, the DIF detection of intercept parameters was nearly 100% accurate except for Item 2 intercept, where non-uniform small DIF was built in.

4) As expected, the power of both slope parameter and intercept parameters increased as the DIF size increased.

Table 14. Power in 20% DIF Item Conditions

Alignment type	Group mean abilities	NG	GS	a2	a4	b1	b2	b3	b4
FIXED	Equal	24	100	0.10	0.12	0.12	0.08	0.19	0.20
			500	0.22	0.63	0.96	0.89	1.00	0.98
		40	100	0.03	0.09	0.30	0.03	0.20	0.43
			500	0.38	0.48	0.99	0.91	0.99	1.00
		80	100	0.02	0.02	0.21	0.02	0.63	0.21
			500	0.20	0.27	0.83	0.90	0.98	0.99
	Unequal	24	100	0.00	0.03	0.34	0.22	0.32	0.03
			500	0.02	0.34	0.77	0.68	0.78	0.98
		40	100	0.01	0.04	0.11	0.13	0.23	0.24
			500	0.11	0.45	0.98	0.90	0.88	1.00
		80	100	0.01	0.04	0.32	0.29	0.61	0.01
			500	0.13	0.47	0.94	0.74	0.94	0.93
FREE	Equal	24	100	0.10	0.08	0.09	0.02	0.23	0.23
			500	0.18	0.57	0.97	0.83	0.98	0.99
		40	100	0.02	0.07	0.24	0.03	0.20	0.48
			500	0.43	0.51	1.00	0.90	1.00	0.98
		80	100	0.00	0.02	0.12	0.03	0.57	0.22
			500	0.22	0.33	0.90	0.78	0.99	1.00
	Unequal	24	100	0.00	0.01	0.11	0.22	0.32	0.02
			500	0.11	0.22	0.79	0.69	0.98	0.89
		40	100	0.00	0.02	0.10	0.10	0.24	0.21
			500	0.12	0.45	0.90	0.89	0.98	1.00
		80	100	0.00	0.02	0.04	0.42	0.41	0.04
			500	0.19	0.42	0.92	0.63	1.00	0.83

The FIXED alignment method tended to be more accurate than the FREE alignment method on detecting DIF from slope parameters, although nearly all powers of slopes are below 0.5. On the intercept parameter, the two alignment methods were equally accurate in conditions with 500 individuals per group, but the FIXED alignment method was slightly more accurate in conditions with 100 individuals per group. This finding is only partially useful for empirical

studies. In conditions where the power is expected to be small, neither estimation method will be considered as reliable, so which one is slightly more accurate on detecting DIF is not of interest. In equal group mean abilities and large group size conditions where the power is expected to be large, both methods had the similar power. Therefore, which method to choose in such conditions will be determined by other criteria, such as Type I error rate.

Table 15 presents the power of all parameters with build-in DIF in conditions with 40% DIF items. All four conclusions summarized from conditions with 20% DIF items remain true in those conditions: 1) The DIF detection on intercept parameters were generally more accurate than the DIF detection on slope parameters within each simulation condition; 2) The DIF detection on all parameters were more accurate in conditions with large group size than in conditions with small group size; 3) The DIF detection was most accurate in conditions with equal group mean abilities than in conditions with unequal group mean abilities; and 4) The power of both slope parameter and intercept parameters increased as the DIF size increased.

The two alignment methods performed similarly in majority of conditions except one. In conditions with unequal group mean abilities and the largest sample size (80 groups × 500 individuals per group), the FREE alignment method tended to be more accurate on detecting DIF from slope parameters, and the FIXED alignment method tended to be more accurate on detecting DIF from intercept parameters. This is less likely a consistent pattern due to certain simulation factors, but more likely because the DIF detection made by both methods were largely influenced by the unequal group mean abilities and led to unreliable results.

A comparison between Table 14 and Table 15 found that the DIF detection on intercept parameters were similarly accurate between conditions with different proportion of DIF items. The

DIF detection on slope parameters where medium size non-uniform DIF exist tended to be more accurate in conditions with larger proportion of DIF items.

Table 15. Power in 40% DIF Item Conditions

Alignment type	Group mean abilities	NG	GS	a2	a4	a6	a8	b1	b2	b3	b4	b5	b6	b7	b8
FIXED	Equal	24	100	0.12	0.19	0.00	0.33	0.19	0.50	0.52	0.30	0.30	0.12	0.64	0.51
			500	0.12	0.39	0.11	0.71	0.98	0.93	0.97	0.92	0.98	0.81	0.98	0.97
		40	100	0.11	0.21	0.01	0.38	0.31	0.11	0.59	0.33	0.20	0.11	0.81	0.51
			500	0.23	0.50	0.30	0.77	1.00	0.96	0.99	0.98	0.91	0.98	0.97	0.99
		80	100	0.09	0.38	0.05	0.23	0.23	0.11	0.72	0.29	0.13	0.21	0.60	0.30
			500	0.14	0.52	0.39	0.72	0.92	0.93	1.00	0.98	0.92	0.84	0.99	0.98
	Unequal	24	100	0.02	0.02	0.11	0.11	0.33	0.32	0.68	0.25	0.43	0.22	0.56	0.23
			500	0.37	0.53	0.13	0.75	0.89	0.49	0.98	0.99	0.98	0.63	1.00	0.64
		40	100	0.00	0.11	0.10	0.10	0.31	0.12	0.84	0.31	0.29	0.22	0.71	0.42
			500	0.11	0.67	0.22	0.78	0.96	0.80	0.99	1.00	0.98	0.99	0.99	0.90
		80	100	0.11	0.01	0.09	0.21	0.34	0.21	0.79	0.44	0.23	0.32	0.45	0.10
			500	0.25	0.58	0.16	0.54	0.99	0.91	1.00	0.93	0.94	0.98	0.93	0.98
FREE	Equal	24	100	0.11	0.16	0.00	0.33	0.13	0.54	0.48	0.32	0.32	0.14	0.63	0.49
			500	0.15	0.39	0.07	0.74	0.98	0.93	0.99	0.82	0.99	0.80	0.99	0.98
		40	100	0.12	0.21	0.00	0.41	0.28	0.12	0.62	0.30	0.19	0.12	0.68	0.50
			500	0.24	0.50	0.28	0.80	0.99	0.98	1.00	0.97	0.91	0.95	1.00	0.99
		80	100	0.01	0.52	0.00	0.15	0.15	0.15	0.63	0.26	0.03	0.13	0.55	0.26
			500	0.13	0.49	0.37	0.76	0.91	0.88	0.98	0.99	0.88	0.89	0.98	1.00
	Unequal	24	100	0.00	0.05	0.12	0.10	0.11	0.10	0.52	0.08	0.34	0.09	0.52	0.18
			500	0.39	0.54	0.14	0.75	0.98	0.54	0.99	1.00	0.99	0.52	1.00	0.88
		40	100	0.00	0.09	0.10	0.13	0.12	0.11	0.83	0.32	0.33	0.21	0.59	0.31
			500	0.12	0.68	0.22	0.76	1.00	0.80	0.97	0.98	0.98	0.90	0.98	0.98
		80	100	0.00	0.04	0.08	0.14	0.43	0.30	0.87	0.28	0.00	0.30	0.43	0.14
			500	0.33	0.83	0.25	0.35	0.93	0.76	0.93	0.84	0.83	0.83	0.93	0.83

Overall, the two alignment methods performed similar in terms of the power of DIF detection on both parameters. The group size and type of group mean abilities made the largest impact on the power of DIF detection, while the proportion of DIF items and number of groups only had limited impact on this. The powers were at largest in conditions with equal group mean abilities and large group size. The type of DIF did not affect the DIF detection on intercept parameters. For items with non-uniform DIF, the power on intercept parameters were much larger than the power on slope parameters.

4.4.2 Type I Error Rate

This section discusses the Type I error rate of the DIF detection on both slope and intercept parameters in different conditions. Table 16 and Table 17 present the Type I error rates in conditions with 20% DIF items. Table 16 demonstrates that the Type I error rates of DIF detection on slope parameter were generally very small, and they decreased as the group size increased. In conditions with 500 individuals per group, the chance of incorrectly identify a non-DIF slope with DIF existing was less than 10%. The largest Type I error rate was in condition with the smallest sample size (24 groups × 100 individuals per group).

Table 16. Type I Error Rate on Slope Parameters in 20% DIF Item Conditions

Alignment type	Group mean abilities	NG	GS	a1	a3	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16	a17	a18	a19	a20
FIXED	Equal	24	100	0.01	0.01	0.00	0.06	0.01	0.02	0.04	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00
FIXED	Equal	24	500	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.00
FIXED	Equal	40	100	0.07	0.06	0.00	0.04	0.00	0.00	0.00	0.01	0.01	0.06	0.03	0.00	0.00	0.00	0.00	0.11	0.07	0.00
FIXED	Equal	40	500	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.00
FIXED	Equal	80	100	0.18	0.01	0.00	0.07	0.00	0.00	0.00	0.04	0.00	0.05	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.01
FIXED	Equal	80	500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.04	0.08	0.02	0.00	0.00	0.00	0.00	0.01	0.00
FIXED	Unequal	24	100	0.00	0.02	0.00	0.03	0.00	0.03	0.05	0.03	0.00	0.22	0.11	0.00	0.09	0.01	0.01	0.02	0.04	0.02
FIXED	Unequal	24	500	0.04	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.03	0.02	0.00	0.05	0.03	0.03	0.00	0.00	0.01
FIXED	Unequal	40	100	0.05	0.08	0.00	0.08	0.02	0.02	0.04	0.06	0.00	0.02	0.00	0.00	0.12	0.10	0.00	0.00	0.00	0.01
FIXED	Unequal	40	500	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.10	0.01	0.00	0.01	0.00	0.00
FIXED	Unequal	80	100	0.12	0.02	0.00	0.01	0.00	0.01	0.00	0.02	0.02	0.01	0.00	0.00	0.07	0.10	0.07	0.02	0.02	0.02
FIXED	Unequal	80	500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.07	0.02	0.01	0.06	0.02	0.03
FREE	Equal	24	100	0.04	0.03	0.00	0.08	0.00	0.00	0.02	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.09	0.02	0.00
FREE	Equal	24	500	0.04	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.00
FREE	Equal	40	100	0.06	0.04	0.00	0.05	0.00	0.00	0.01	0.03	0.01	0.06	0.04	0.01	0.00	0.03	0.03	0.08	0.06	0.00
FREE	Equal	40	500	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00
FREE	Equal	80	100	0.22	0.01	0.00	0.10	0.01	0.00	0.03	0.07	0.01	0.11	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.01
FREE	Equal	80	500	0.02	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.10	0.10	0.00	0.00	0.00	0.00	0.01	0.00	0.00
FREE	Unequal	24	100	0.02	0.02	0.00	0.03	0.00	0.00	0.09	0.04	0.03	0.23	0.12	0.00	0.10	0.04	0.01	0.02	0.01	0.02
FREE	Unequal	24	500	0.05	0.00	0.00	0.01	0.00	0.01	0.02	0.10	0.00	0.03	0.05	0.00	0.01	0.01	0.02	0.00	0.02	0.00
FREE	Unequal	40	100	0.13	0.11	0.00	0.11	0.03	0.03	0.03	0.08	0.00	0.01	0.03	0.00	0.10	0.09	0.00	0.00	0.03	0.01
FREE	Unequal	40	500	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.00	0.01	0.01	0.00	0.11	0.00	0.00	0.00	0.00	0.03

FREE	Unequal	80	100	0.03	0.01	0.00	0.02	0.00	0.16	0.04	0.04	0.01	0.02	0.01	0.01	0.02	0.15	0.02	0.03	0.01	0.00
FREE	Unequal	80	500	0.01	0.00	0.00	0.00	0.00	0.03	0.00	0.02	0.02	0.00	0.00	0.02	0.09	0.02	0.00	0.07	0.01	0.00

The DIF detection on intercepts in conditions with 20% DIF items was generally accurate with Type I error rate smaller than 0.1. However, Table 16 shows two unique patterns that have not been found before.

1) Increase the group size would not necessarily lead to a smaller Type I error rate. In many cases, the Type I error rates were larger in conditions with 500 individuals per group than in conditions with 100 individuals per group, while other factors were controlled.
2) The Type I error rate on intercepts were at largest when FIXED alignment method was used for estimation on datasets with unequal group mean abilities. In this condition, the Type I error rates consistently increased as the group size increased, and they were equally large across different number of groups. According to Section 4.4.1, the two alignment methods performed similarly across conditions in terms of the power. This finding suggested that the FREE alignment method should be preferred over the FIXED alignment method in empirical studies with unequal group mean abilities to avoid large Type I error rate.

Table 17. Type I Error Rate on Intercept Parameters in 20% DIF Item Conditions

Alignment type	Group mean abilities	NG	GS	b5	b6	b7	b8	b9	b10	b11	b12	b13	b14	b15	b16	b17	b18	b19	b20
FIXED	Equal	24	100	0.02	0.00	0.07	0.00	0.01	0.00	0.02	0.01	0.04	0.00	0.03	0.00	0.01	0.00	0.01	0.00
FIXED	Equal	24	500	0.00	0.01	0.01	0.01	0.07	0.00	0.00	0.00	0.15	0.00	0.02	0.01	0.03	0.02	0.02	0.00
FIXED	Equal	40	100	0.00	0.02	0.02	0.00	0.01	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.04	0.01	0.03	0.03
FIXED	Equal	40	500	0.00	0.00	0.00	0.00	0.00	0.03	0.05	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.01
FIXED	Equal	80	100	0.01	0.00	0.01	0.02	0.00	0.02	0.01	0.01	0.08	0.00	0.01	0.04	0.00	0.00	0.01	0.00
FIXED	Equal	80	500	0.03	0.00	0.00	0.00	0.02	0.00	0.02	0.03	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.00
FIXED	Unequal	24	100	0.00	0.10	0.11	0.10	0.11	0.22	0.10	0.02	0.10	0.02	0.00	0.21	0.00	0.03	0.06	0.02
FIXED	Unequal	24	500	0.22	0.22	0.21	0.32	0.22	0.23	0.01	0.34	0.11	0.43	0.23	0.11	0.01	0.06	0.07	0.21
FIXED	Unequal	40	100	0.02	0.04	0.02	0.12	0.10	0.04	0.03	0.13	0.09	0.01	0.10	0.01	0.02	0.01	0.00	0.02
FIXED	Unequal	40	500	0.10	0.11	0.00	0.01	0.09	0.10	0.00	0.11	0.01	0.11	0.09	0.10	0.00	0.10	0.04	0.13

FIXED	Unequal	80	100	0.08	0.09	0.08	0.09	0.02	0.01	0.08	0.01	0.02	0.10	0.04	0.10	0.06	0.02	0.02	0.04
FIXED	Unequal	80	500	0.13	0.12	0.07	0.07	0.34	0.26	0.27	0.06	0.14	0.17	0.12	0.16	0.14	0.26	0.10	0.15
FREE	Equal	24	100	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.01	0.00	0.00	0.00	0.01
FREE	Equal	24	500	0.00	0.00	0.01	0.00	0.06	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.03	0.01	0.00
FREE	Equal	40	100	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.02	0.01	0.02	0.00	0.00	0.01	0.02	0.00
FREE	Equal	40	500	0.00	0.00	0.00	0.01	0.05	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
FREE	Equal	80	100	0.02	0.00	0.00	0.01	0.00	0.08	0.01	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.01
FREE	Equal	80	500	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.02	0.00	0.02
FREE	Unequal	24	100	0.00	0.02	0.01	0.00	0.04	0.00	0.01	0.02	0.02	0.02	0.00	0.01	0.01	0.00	0.03	0.03
FREE	Unequal	24	500	0.00	0.01	0.11	0.00	0.01	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.01	0.02	0.01
FREE	Unequal	40	100	0.00	0.00	0.01	0.02	0.08	0.00	0.00	0.10	0.08	0.00	0.00	0.00	0.01	0.01	0.02	0.00
FREE	Unequal	40	500	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.00	0.00
FREE	Unequal	80	100	0.03	0.00	0.00	0.00	0.01	0.01	0.02	0.02	0.01	0.03	0.03	0.02	0.00	0.01	0.00	0.02
FREE	Unequal	80	500	0.01	0.00	0.02	0.00	0.00	0.02	0.01	0.01	0.02	0.00	0.00	0.01	0.00	0.02	0.01	0.01

Table 18 and Table 19 below present the Type I error rates on each parameter in conditions with 40% DIF items. In general, the Type I error rates on slopes in conditions with 40% DIF items were slightly larger than those in conditions with 20% DIF items, but the general pattern remains the same. The overall Type I error rates were small, and they decreased as the group size increased. The Type I error rate on slopes tended to be larger on items with small non-uniform DIF. Among those relatively large Type I error rates, the largest values come from conditions with unequal group mean abilities and small group size, regardless of estimation methods.

Table 18. Type I Error Rate on Slope Parameters in 40% DIF Item Conditions

Alignment type	Group mean abilities	NG	GS	a1	a3	a5	a7	a9	a10	a11	a12	a13	a14	a15	a16	a17	a18	a19	a20
FIXED	Equal	24	100	0.01	0.01	0.00	0.04	0.06	0.05	0.00	0.10	0.00	0.04	0.01	0.00	0.01	0.00	0.05	0.02
FIXED	Equal	24	500	0.10	0.00	0.00	0.00	0.01	0.01	0.00	0.08	0.00	0.07	0.02	0.00	0.00	0.00	0.01	0.00
FIXED	Equal	40	100	0.12	0.00	0.02	0.02	0.08	0.04	0.03	0.01	0.02	0.00	0.02	0.00	0.20	0.02	0.00	0.04
FIXED	Equal	40	500	0.03	0.00	0.00	0.00	0.02	0.02	0.00	0.03	0.00	0.00	0.06	0.00	0.19	0.00	0.01	0.01
FIXED	Equal	80	100	0.02	0.00	0.00	0.06	0.02	0.01	0.00	0.04	0.04	0.00	0.03	0.00	0.00	0.00	0.07	0.03
FIXED	Equal	80	500	0.08	0.00	0.00	0.01	0.01	0.01	0.00	0.05	0.08	0.00	0.00	0.00	0.02	0.01	0.08	0.08
FIXED	Unequal	24	100	0.03	0.13	0.01	0.00	0.03	0.03	0.01	0.00	0.01	0.02	0.21	0.01	0.10	0.10	0.11	0.01
FIXED	Unequal	24	500	0.01	0.25	0.00	0.00	0.00	0.01	0.00	0.01	0.02	0.01	0.04	0.01	0.12	0.02	0.12	0.00
FIXED	Unequal	40	100	0.02	0.09	0.00	0.06	0.04	0.20	0.00	0.02	0.21	0.15	0.00	0.00	0.01	0.01	0.18	0.00
FIXED	Unequal	40	500	0.05	0.00	0.00	0.09	0.02	0.02	0.00	0.00	0.02	0.00	0.01	0.00	0.11	0.02	0.10	0.01

FIXED	Unequal	80	100	0.45	0.03	0.22	0.11	0.10	0.00	0.04	0.00	0.04	0.03	0.02	0.02	0.00	0.11	0.09	0.03
FIXED	Unequal	80	500	0.08	0.00	0.09	0.08	0.07	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.02	0.00	0.02	0.00
FREE	Equal	24	100	0.04	0.01	0.00	0.01	0.10	0.06	0.00	0.09	0.00	0.04	0.00	0.00	0.00	0.00	0.05	0.00
FREE	Equal	24	500	0.08	0.00	0.00	0.07	0.03	0.01	0.00	0.17	0.04	0.06	0.00	0.00	0.00	0.00	0.00	0.01
FREE	Equal	40	100	0.09	0.02	0.04	0.01	0.07	0.07	0.00	0.01	0.00	0.00	0.03	0.00	0.22	0.03	0.02	0.10
FREE	Equal	40	500	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.20	0.00	0.01	0.02
FREE	Equal	80	100	0.02	0.00	0.00	0.13	0.02	0.03	0.00	0.00	0.12	0.00	0.01	0.00	0.01	0.00	0.01	0.00
FREE	Equal	80	500	0.12	0.00	0.00	0.02	0.01	0.01	0.00	0.00	0.11	0.02	0.00	0.00	0.00	0.00	0.00	0.13
FREE	Unequal	24	100	0.02	0.10	0.00	0.01	0.01	0.02	0.02	0.00	0.03	0.00	0.16	0.02	0.08	0.06	0.20	0.01
FREE	Unequal	24	500	0.02	0.24	0.00	0.00	0.02	0.00	0.04	0.00	0.00	0.00	0.02	0.00	0.12	0.01	0.11	0.00
FREE	Unequal	40	100	0.05	0.08	0.02	0.03	0.04	0.20	0.00	0.02	0.16	0.18	0.01	0.00	0.04	0.02	0.21	0.02
FREE	Unequal	40	500	0.03	0.00	0.00	0.09	0.01	0.02	0.00	0.00	0.01	0.02	0.00	0.00	0.09	0.00	0.12	0.00
FREE	Unequal	80	100	0.42	0.05	0.28	0.02	0.14	0.01	0.01	0.00	0.02	0.01	0.00	0.01	0.01	0.14	0.01	0.00
FREE	Unequal	80	500	0.09	0.00	0.08	0.00	0.07	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.01

Consistent with the finding from slope parameters, the Type I error rates on intercepts in conditions with 40% DIF items were slightly larger than those in conditions with 20% DIF items, but the overall Type I error rates were still small. In general, the Type I error rates on intercepts increased as the group size increased. When group mean abilities were unequal and the FIXED alignment method was used for parameter estimation, the Type I error rates were clearly larger than the average Type I error rates from other conditions, and this inaccuracy increased as the group size increased.

Table 19. Type I Error Rate on Intercept Parameters in 40% DIF Item Conditions

Alignment type	Group mean abilities	NG	GS	b9	b10	b11	b12	b13	b14	b15	b16	b17	b18	b19	b20
FIXED	Equal	24	100	0.00	0.00	0.01	0.00	0.01	0.08	0.01	0.00	0.08	0.01	0.00	0.01
FIXED	Equal	24	500	0.00	0.00	0.00	0.01	0.00	0.09	0.02	0.00	0.00	0.00	0.02	0.06
FIXED	Equal	40	100	0.00	0.00	0.12	0.04	0.10	0.00	0.03	0.01	0.03	0.05	0.00	0.02
FIXED	Equal	40	500	0.01	0.02	0.00	0.00	0.08	0.00	0.07	0.00	0.02	0.00	0.03	0.00
FIXED	Equal	80	100	0.00	0.00	0.00	0.18	0.02	0.02	0.01	0.00	0.07	0.03	0.08	0.00
FIXED	Equal	80	500	0.00	0.00	0.00	0.00	0.09	0.01	0.00	0.00	0.01	0.10	0.01	0.01
FIXED	Unequal	24	100	0.12	0.09	0.02	0.02	0.03	0.11	0.05	0.02	0.11	0.11	0.12	0.00
FIXED	Unequal	24	500	0.25	0.24	0.00	0.00	0.24	0.12	0.12	0.01	0.01	0.02	0.00	0.12
FIXED	Unequal	40	100	0.01	0.08	0.30	0.07	0.09	0.08	0.11	0.06	0.02	0.15	0.02	0.03

FIXED	Unequal	40	500	0.11	0.10	0.12	0.22	0.11	0.13	0.04	0.01	0.00	0.22	0.02	0.24
FIXED	Unequal	80	100	0.10	0.03	0.11	0.22	0.02	0.01	0.10	0.07	0.03	0.12	0.07	0.01
FIXED	Unequal	80	500	0.42	0.53	0.32	0.43	0.24	0.23	0.32	0.42	0.25	0.31	0.43	0.42
FREE	Equal	24	100	0.00	0.00	0.00	0.00	0.04	0.06	0.00	0.01	0.02	0.00	0.00	0.00
FREE	Equal	24	500	0.00	0.01	0.00	0.00	0.00	0.10	0.00	0.00	0.01	0.01	0.00	0.05
FREE	Equal	40	100	0.00	0.00	0.07	0.00	0.02	0.00	0.01	0.03	0.01	0.04	0.00	0.01
FREE	Equal	40	500	0.00	0.00	0.01	0.00	0.06	0.01	0.04	0.01	0.00	0.02	0.01	0.01
FREE	Equal	80	100	0.00	0.00	0.00	0.12	0.01	0.00	0.00	0.00	0.00	0.00	0.04	0.03
FREE	Equal	80	500	0.01	0.00	0.00	0.02	0.12	0.01	0.00	0.00	0.00	0.01	0.00	0.00
FREE	Unequal	24	100	0.00	0.03	0.04	0.00	0.01	0.05	0.02	0.00	0.00	0.03	0.02	0.01
FREE	Unequal	24	500	0.00	0.01	0.01	0.01	0.14	0.02	0.03	0.00	0.01	0.00	0.03	0.11
FREE	Unequal	40	100	0.00	0.08	0.10	0.05	0.11	0.01	0.00	0.02	0.03	0.02	0.00	0.01
FREE	Unequal	40	500	0.00	0.00	0.03	0.02	0.02	0.00	0.01	0.01	0.00	0.06	0.00	0.02
FREE	Unequal	80	100	0.02	0.01	0.14	0.01	0.02	0.01	0.14	0.00	0.00	0.14	0.02	0.04
FREE	Unequal	80	500	0.00	0.07	0.08	0.01	0.01	0.02	0.03	0.01	0.01	0.01	0.08	0.00

Overall, the DIF detection on both slope and intercept parameters were accurate with small Type I error rates (< .2) in majority of conditions. The Type I error rates on both parameters tended to be larger in conditions with larger proportion of DIF items. The number of groups made no impact on the Type I error rate. The impact of group size was not as strong as for powers, and this impact was not consistent between slope and intercept parameters. When small non-uniform DIF existed, the Type I error rates on the corresponding slopes tended to be larger. The impact of group mean abilities on the Type I error rates on slopes was limited. However, FIXED alignment method used in conditions with unequal group mean abilities would result in clearly larger Type I error rates (> .4) on intercepts. In conclusion, the FIXED alignment method should be avoided for datasets with unequal group mean abilities if DIF detection is of interest.

4.4.3 Summary

Section 4.4.1 and 4.4.2 have proved that group mean abilities and group size are the two most important influential factors for DIF detection. Table 20 summarized the results by excluding

proportion of DIF and number of groups, and averaging the powers and Type I error rates across items and conditions. The following conclusions can be drawn from this summary table.

Table 20. Summary of DIF Detection Accuracy

Alignment method	Group mean abilities	Group size	Power on slopes	Power on intercepts	Type I error rate on slopes	Type I error rate on intercepts
FIXED	Equal	100	0.12	0.29	0.02	0.02
FIXED	Equal	500	0.39	0.95	0.01	0.02
FIXED	Unequal	100	0.05	0.31	0.05	0.07
FIXED	Unequal	500	0.34	0.89	0.02	0.16
FREE	Equal	100	0.11	0.27	0.03	0.01
FREE	Equal	500	0.39	0.95	0.02	0.01
FREE	Unequal	100	0.04	0.25	0.05	0.02
FREE	Unequal	500	0.35	0.88	0.02	0.02

The DIF detection on slope parameter was not reliable with small powers, regardless of estimation method. The DIF detection in conditions with small group size (GS = 100) was not reliable with small powers, regardless of item parameter and estimation method. The DIF detection on intercept parameter in conditions with equal group mean abilities and large group size was highly reliable with large powers and small Type I error rates. The DIF detection made by the FREE alignment method on intercept parameter in conditions with unequal group mean abilities and large group size was fairly reliable with relatively large powers and small Type I error rates. However, the DIF detection made by the FIXED alignment method in such conditions was not reliable due to the inflated Type I error rates.

5.0 Discussion

5.1 Findings from the Simulation Study

5.1.1 Comparison Between Estimation Methods

In this study, all three estimation methods were accurate on parameter estimation in majority of simulation conditions. The two alignment methods tended to be more accurate than the random item effects model on the overall parameter recovery after controlling for other factors. The overall mean RMSD of expected score made by the random item effects model, the FIXED alignment method, and the FREE alignment method was 0.21, 0.34, and 0.44, respectively.

In terms of the individual parameter, the random item effects model was more accurate on estimating slope parameters, and the two alignment methods were more accurate on estimating intercept parameters. The alignment methods tended to overestimate both slope and intercept parameters, while the random item effects model tended to overestimate slope parameters and underestimate intercept parameters. The random item effects model was accurate on estimating the slope parameter with mean bias equaled to 0.02 and mean RMSD equaled to 0.12. If appropriate DIF analysis method was followed, high accuracy of DIF detection on the slope parameter can be achieved. However, the estimation of slope parameter made by the random item effects model could be highly in accurate in conditions with unequal group mean abilities and small group size. The impact of different factors on slope estimation will be discussed in the next section. The alignment method was more accurate on estimating intercept parameters after controlling for other factors. The mean bias and mean RMSD of intercept parameter made by the

two alignment methods was -0.3 and 0.29, respectively. The estimates of intercept made by the random item effects was not only less accurate than those made by the alignment method, but also tended to be inconsistent across individual items.

The two alignment methods performed similarly on parameter estimation. In many conditions, the FREE alignment method was slightly more accurate than the FIXED alignment method with a small significant difference. Although the difference between the two alignment methods tended to be small statistically, the consistent advantage of the FREE alignment method makes it a preferred method between those two. This finding confirmed Muthén and Asparouhov (2014)'s statement that the FREE alignment method is more accurate than the FIXED alignment method.

The alignment method can flag slope and intercept parameters with DIF directly. This study found that the DIF detection on intercept parameters could be highly accurate with large power and small Type I error rate in many conditions. As a result of the similarity on parameter estimation, the two alignment methods performed similarly on DIF detection. The specific differences between them will be discussed in the next section. Overall, the FREE alignment method is recommended over the FIXED alignment method. The FIXED alignment method was unreliable on estimating intercept parameters in conditions with unequal group mean abilities and small group size. Therefore, it should not be used if the estimate of intercept or the DIF in intercept parameter is of interest.

5.1.2 Effect of Different Factors

Traditional DIF analysis methods are only applicable between two groups or across a handful of groups where pairwise comparison can be made. One of the most essential research

questions in this study was how the random item effects model and the alignment methods performed in conditions with more than five groups. Previous studies have already proved that the alignment method can make unbiased parameter estimates on datasets with 2 to 60 groups, but with much larger group sizes (Asparouhov, 2014; Finch, 2016). This study found that the group size played a more critical role than the number of groups on both parameter estimation and on DIF detection. The number of groups made very limited impact on the overall parameter recovery while other factors were controlled. It indicates that all three methods can perform relatively consistent across a variety number of groups. On the individual item level, the number of groups had a small significant impact on the estimation of slopes and intercepts. It has been found that the random item effects model performed the worst on estimating slope parameters in conditions with insufficient level-2 sample size (NG = 24). Once the level-2 sample size was large enough (NG = 40), increased the numbers of groups would not increase the estimation accuracy on slope parameter. The number of groups made no impact on the estimation of intercept parameter made by the random item effect model, but it had a small significant impact on the estimation of intercept made by the two alignment methods, especially those made by the FIXED alignment method. The intercept parameter estimation accuracy made by the two alignment methods tended to gradually decrease as the number of groups decreases. While the two alignment methods performed similar in conditions with 80 groups, the FIXED alignment method was significantly less accurate than the FREE alignment method in conditions with 24 groups. Overall, none of the differences discussed above has a large enough effect size to confidently generalize the results to empirical studies. It can be concluded that tthe performance of all methods was generally consistent across conditions with 24 group, 40 groups, and 80 groups, despite the small significant differences detected from some conditions. This encouraging finding demonstrated that while other factors are

controlled, any of those three methods can perform equally well across varying large numbers of groups.

The group size had a strong significant impact on the parameter estimation and DIF detection. All methods were fairly accurate on parameter recovery with small biases and RMSDs in conditions with large group size (GS = 500). The accuracy of parameter estimation made by all three methods decreased in conditions with small group size (GS = 100), and this inaccuracy cannot be compromised by increasing the number of groups. Five-hundred or more individuals per group has been proved as a sufficient group size for unbiased parameter recovery by many previous studies (Asparouhov & Muthén, 2014; De Boeck, 2008; Stark et al., 2006), but only a few studies (MacCallum et al., 1999; Meade & Lautenschlager, 2004) explored a minimum sample size below 500. Group size as small as 100 or 150 may be enough for unbiased parameter estimates, if other factors were carefully controlled, such as communalities around 0.5 or invariant factor loadings. This study found small group size would amplify the impact of unequal group mean abilities. More specifically, parameter estimates in such conditions were often heavily biased. As the group size increased, the estimation accuracy of slope parameters made by all three methods increased, but those made by the alignment methods changed more dramatically than those made by the random item effects model. In other words, the slope parameter estimates made by the alignment methods in conditions with 100 individuals per group were highly inaccurate and useless, while the slope parameter estimates made by the random item effects model in the same condition can still lead to reasonably reliable results. The DIF detection of slope parameters in conditions with 500 individuals per group were generally accurate with large power and small Type I error rate, and this accuracy dropped largely with smaller power and inflated Type I error rate in conditions with 100 individuals per group.

The type of group mean abilities is another factor that made strong significant impact on parameter estimation and DIF detection, especially on the intercept parameters. This effect impacted both the random item effects model and the alignment method, which is consistent with De Boeck (2008)'s finding from the RIP model and Finch (2016)'s finding from the alignment method. The estimation accuracy of intercept parameter made by the two alignment methods decreased dramatically as the group mean abilities became unequal, and the estimation made by the FIXED alignment method was more likely to be influenced by the unequal group mean abilities. The steeply decreased estimation accuracy leaded to inaccurate DIF detection of intercept parameters. When FIXED alignment method was used in conditions with unequal group mean abilities, although the power of DIF detection on intercepts remained large, the Type I error rate became large as well. The DIF detection on intercept parameter was generally accurate except when FIXED alignment method was used in conditions with unequal group mean abilities. One step further than previous studies, this study evaluated the accuracy of DIF detection in conditions with unequal group mean abilities and small DIF size. The result shows that the power on slopes tended to be very small in such conditions, and the Type I error rate on intercepts may be inflated.

Three other factors were also investigated in this study. The proportion of DIF items only made a limited impact on the parameter estimation and DIF detection. The DIF detection of intercepts were not influenced by the type of DIF. As expected, the larger the DIF size was, the more accurate DIF detection results would be.

In summary, the random item effects model performed differently with the alignment methods in majority of simulation conditions. On the individual parameter level, the alignment methods tended to be more accurate on estimating intercepts, while the random item effects model tended to be more accurate on estimating slopes. The alignment methods tended to be more

accurate on the overall parameter recovery and were preferred in many simulation conditions with two exceptions. 1) The alignment method, especially the FIXED alignment method, may lead to inaccurate parameter estimates and inflated Type I error rate on the intercept parameter in conditions with unequal group mean abilities. 2) The random item effects model was significantly more accurate on estimating slopes in conditions with small group size than the alignment methods. The inaccurate estimates of slope made by the alignment methods would lead to inaccurate DIF detection on slopes. This finding suggests that the random item model should be used if the slope parameter is of interest, despite the many other advantages of the alignment methods.

In terms of the computational efficiency, the difference in estimation time between three methods are not comparable with the difference in time caused by the increasing sample size and the overall convergence. In general, larger datasets would take longer to converge. A dataset with homogeneous dichotomous responses would increase the number of iterations significantly, and sometimes convergence could not be achieved even with 50,000 iterations. This study found that when the required number of iterations was large (> 5,000), there was a chance to detect poor trace plots and poor autocorrelation plot even the model converged in Mplus, and this chance increased as the required number of iterations increased. In empirical studies, if the convergence is fast and valid, which estimation to use would not make much difference regarding the estimation time.

5.2 Recommendations for Empirical Studies

It needs to be clarified that the recommendations listed below were proposed based on findings from this study, which means they can only be applied to datasets with dichotomous outcome variables.

The two alignment methods are generally more accurate and more efficient as compared to the random item effects model. They are accurate on the overall parameter recovery and on estimation of intercept parameters. They can directly flag item parameters that have DIF existing, and this DIF identification can be highly accurate in conditions with large group size and equal group mean abilities. In addition to the Bayesian estimation, they also work with maximum likelihood estimation, which could increase the computational efficiency greatly while sacrificing the estimation accuracy to some extent. In general, the alignment methods are recommended over the random item effects model regardless of the number of groups, proportion of DIF items, DIF type, and DIF size. In addition, the alignment methods have a small significant advantage on the overall parameter recovery than the random item effects model in conditions with 24 groups as compared to the discrepancies in conditions with 40 and 80 groups. This finding is consistent with Muthén and Asparouhov (2013)'s conclusion that the alignment method is suggested for data with less than 30 groups.

In this study, the two alignment methods performed similarly with no significant difference or only a small significant difference between them in many conditions. In conditions where notable significant differences existed, the FREE alignment method was always the one with higher parameter estimation accuracy and higher DIF detection accuracy. This finding is consistent with Muthén and Asparouhov (2014)'s conclusion. Therefore, the FREE alignment method is

generally recommended over the FIXED alignment method unless there are only two groups, where the FREE alignment method is not applicable.

The FREE alignment method is highly accurate on detecting DIF in intercept parameters in conditions with large group size. The FIXED alignment method is also accurate in the same conditions, unless the group mean abilities are unequal. Neither of the two alignments are accurate on detecting DIF in slope parameters. When the group size is large, a moderate power of DIF detection on slopes can sometimes be achieved with medium size non-uniform DIF. But in empirical studies, there is no way to know what the true DIF type and true DIF size are. Therefore, the alignment methods are simply not recommended for detecting DIF in slope parameters. According to Equation (6) and (7), the slope parameter from the CFA model can be converted to the item discrimination parameter from the 2PL IRT model, and the intercept parameter can be converted to the product of item difficulty and item discrimination parameters. It indicates that the alignment methods can identify items with DIF directly with high accuracy, but it cannot distinguish whether this DIF is uniform or non-uniform. The alignment methods are recommended as the initial step of analysis in empirical studies. They are straightforward and efficient. A single set of command can analyze all items simultaneously. Once the items with DIF are identified, other traditional DIF analysis methods can be followed to further evaluate those items.

Although the alignment methods have many advantages, they are less preferred or should be avoided in the following situations:

1) The random item effects model is recommended when the DIF in slope parameter is of interest. Even though the alignment methods are more accurate on the overall parameter estimation and intercept parameter estimation, the random item effects model is more accurate on estimating slope parameters. In addition, it has been proved that the DIF

detection on slope parameter made by the alignment methods are inaccurate and unreliable. In empirical studies, the random item effects model should be used if the DIF in slope parameter/item discrimination parameter is of interest. Once accurate slope parameter estimates are obtained, traditional DIF analysis methods can be adapted to detect DIF using those parameter estimates.

2) None of the methods is reliable in conditions with unequal group mean abilities and small level-1 sample size. The random item effects model is likely to generate highly inaccurate estimates of slope in those conditions. The alignment methods also tend to be less accurate on estimating slope and intercept parameters in those conditions, and they would lead to incorrect DIF detection results. When having such dataset, the data must be adjusted before applying any of those three methods.
3) The FIXED alignment method is generally less preferred than the FREE alignment method. More importantly, the FIXED alignment method should not be used when there is a large variance among group mean abilities. Although the power of DIF detection of intercepts may still be large, the Type I error rate will be inflated, which will lead to unreliable DIF detection results.

5.3 Limitations

5.3.1 Minimum Group Size

This study found that the alignment methods could be inaccurate and are not recommended for dataset with less than 100 individuals per group. It is important to further explore the level-1

sample size and find out what is the minimum requirement for an accurate parameter recovery and DIF detection. One hundred individuals per group has been demonstrated large enough for dataset with ordinary outcome variables or if the communalities of factors are around 0.5. This study confirmed that 100 sample size for dichotomous outcome variables will result in highly inaccurate parameter estimates. Also, small group size will amplify the impact of unequal group mean abilities and create even larger biases in parameter estimates.

In empirical studies, some of the most common question formats, such as True/False question, multiple choices question, etc., all have dichotomous outcomes. Therefore, it is critical to know the threshold of level-1 sample size for dichotomous outcomes. Level-1 sample size is less an issue for cross-national assessments. However, understanding the minimum requirement of level-1 sample size for dataset with dichotomous responses will allow the alignment method to be used in a broader context. It has been demonstrated that alignment methods have great advantages over the random item effects model in conditions with relatively small number of groups (NG = 24). Muthén and Asparouhov (2013) also pointed out that the alignment methods are preferred in conditions with 0 to 30 groups. It would be interesting to know how the alignment method would perform in conditions with smaller number of groups, and what is the minimum group size required in such conditions.

5.3.2 Variance of Group Mean Abilities

This study found that the random item effects model was not affected by the type of group mean abilities, but the unequal group mean abilities reduced the parameter estimation accuracy made by the alignment methods significantly.

This study used two steps to generate conditions with unequal group mean abilities. First, the individual ability parameter θ_{jg} are generated from a normal distribution $\mathrm{N}(\beta_g, 1)$ with a group mean of β_g and a group variance of 1. Second, the group mean ability β_g is generated from a normal distribution with a mean of 0 and a variance of 1, $\mathrm{N}(0, 1)$. The overall ICC was around 0.5, which is a common value in cross-national assessments. However, the 0.5 ICC is still a relatively vague value. This study did not define "extreme" individual abilities or "extreme" group mean abilities, and it did not exclude any extreme values from the analysis either. When a large proportion of parameter estimates were accurate, the overall biases of RMSDs would be largely impact by the inaccurate estimates resulted by extreme true values.

More analyses can be done to investigate the impact of unequal group mean abilities by carefully design the distribution of individual abilities. For example, the extremely large or small group mean abilities could be excluded at the data generation step, the distribution of group mean abilities may not necessarily be normal. Analyses can be done to explore how "equal" the group mean abilities need to be to achieve accurate parameter estimates. More importantly, how "equal" the group mean abilities need to be to achieve accurate parameter estimates in conditions with small level-1 sample size.

5.3.3 Confounding DIF Type and DIF Size

In this study, the true parameter values for each item were randomly generated. As a result, the DIF size and DIF type on individual parameters would confound with the true parameter values. Although patterns about the accuracy of DIF detection could still be found, no statistical conclusion can be drawn regarding the effect of DIF size and DIF type on the accuracy of parameter estimation.

In order to provide strong statistical evidence for the impact of DIF size and DIF type, a dedicated study with multiple datasets sharing the same set of true values is necessary. All confounding factors need to be carefully controlled, then different DIF values can be added to only one item in each dataset. It has been proved that larger DIF size would lead to higher DIF detection accuracy, but it would be interesting to know the ratio between the change in DIF size and the gain of loss in parameter estimation accuracy.

Appendix A

Data Generation and Analysis Code

Data Generation SAS

A.1.1 Conditions with 20% DIF Items

```
/*----------set parameters----------*/
%let nr=100;
%let ng1=24; %let ng2=40; %let ng3=80;
%let gs1=100; %let gs2=500;

/*---------data generation---------*/
%macro generate4;

%do rep=1 %to &nr;

/*----------parameter matrix----------*/
%let seed1=(400000+&rep*10000+1);
%let seed2=(400000+&rep*10000+2);

proc iml;    *a parameter;
  call randseed(&seed1); *&seed1;
  temp_4a0=j(1,20,.);
  call randgen(temp_4a0,'lognormal',0,.3);
  a0=repeat(temp_4a0,4,1);
  use simu.dif4_a; *read in the dif values;
    read all var _all_ into difa;
  close simu.dif4_a;
  a=a0+difa;
  create dif4_a_item1 from a[colname={"a1" "a2" "a3" "a4" "a5" "a6" "a7" "a8"
"a9"
     "a10" "a11" "a12" "a13" "a14" "a15" "a16" "a17" "a18" "a19" "a20"}];
  append from a;
quit;
proc iml;    *b parameter;
  call randseed(&seed2); *&seed2;
  temp_4b0=j(1,20,.);
  call randgen(temp_4b0,'normal',0,.5);
  b0=repeat(temp_4b0,4,1);
  use simu.dif4_b;
```

```
    read all var _all_ into difb;
  close simu.dif4_b;
  b=b0+difb;
  create dif4_b_item1 from b[colname={"b1" "b2" "b3" "b4" "b5" "b6" "b7" "b8"
"b9"
     "b10" "b11" "b12" "b13" "b14" "b15" "b16" "b17" "b18" "b19" "b20"}];
  append from b;
quit;

/*----combine a and b parameter, run group replications----*/
%do f1=1 %to 3;       *factor 1, group replication;
  %if &f1=1 %then %let grep=&ng1/4;
  %if &f1=2 %then %let grep=&ng2/4;
  %if &f1=3 %then %let grep=&ng3/4;

%do grouprep=1 %to &grep;

data dif4_combine1;
  merge dif4_a_item1 dif4_b_item1;
  group0=_n_;
  group=group0+(&grouprep-1)*4;
run;

/*---------individual ability--------*/
%do f2=1 %to 2;       *factor 2, group size;
  %if &f2=1 %then %let gs=&gs1;
  %if &f2=2 %then %let gs=&gs2;

%let seed3=(400000+&rep*10000+&f1*1000+&f2*100+3);
%let seed4=(400000+&rep*10000+&f1*1000+&f2*100+4);

proc iml;
  call randseed(&seed3);
  theta_equal=j(&gs,4,.);
  beta=j(&gs,4,.);
  theta_unequal=j(&gs,4,.);
  call randgen(theta_equal,'normal',0,1);
  call randgen(beta,'normal',0,1);
  call randgen(theta_unequal,'normal',beta,1);
  create dif4equal_ind1 from theta_equal[colname={"g1" "g2" "g3" "g4"}];
  append from theta_equal;
  create dif4unequal_ind1 from theta_unequal[colname={"g1" "g2" "g3" "g4"}];
  append from theta_unequal;
quit;
data dif4equal_ind2;
  set dif4equal_ind1;
  person=_n_; meanab=1;  *mean ability equal;
  array g(4) g1-g4;
  do group0=1 to 4;
    theta=g(group0);  *theta in equal ability condition;
      output;
  end;
run;
data dif4unequal_ind2;
  set dif4unequal_ind1;
  person=_n_; meanab=2;  *mean ability unequal;
  array g(4) g1-g4;
```

```
  do group0=1 to 4;
    theta=g(group0);  *theta in unequal ability condition;
      output;
  end;
run;
proc sort data=dif4equal_ind2;
  by group0; run;
data dif4equal_combine2;
  merge dif4equal_ind2(drop=g1 g2 g3 g4) dif4_combine1;
  by group0;
run;
proc sort data=dif4unequal_ind2;
  by group0; run;
data dif4unequal_combine2;
  merge dif4unequal_ind2(drop=g1 g2 g3 g4) dif4_combine1;
  by group0;
run;
data dif4_combine3;
  set dif4equal_combine2 dif4unequal_combine2;
run;

/*-----------item response------------*/
data dif4_combine4;
  set dif4_combine3;
  call streaminit(&seed4);
    array p(20) p1-p20;
      array r(20) r1-r20;
      array item(20) item1-item20;
      array a(20) a1-a20;
      array b(20) b1-b20;
        do i=1 to 20;
          p(i)=probnorm(a(i)*theta-b(i));
            r(i)=rand('uniform',0,1);
            if p(i)>=r(i) then item(i)=1;
            else item(i)=0;
        end;
run;
data dif4_combine5;
  set dif4_combine4 (keep=group person theta a1-a20 b1-b20 p1-p20 item1-
item20 meanab);
  rep=&rep;
  dif=4;
run;
proc sort data=dif4_combine5;
  by rep meanab group person;
run;
data dif4_combine6;
  retain dif rep meanab group person;
  set dif4_combine5;
run;
/*-------data generation done---------*/

data dif4equal dif4unequal;
  set dif4_combine6;
  if meanab=1 then output dif4equal;
  if meanab=2 then output dif4unequal;
run;
```

```
%let cond1=%eval(4000+100+&f1*10+&f2*1);     *100 indicate equal mean ability
condition;
%let cond2=%eval(4000+200+&f1*10+&f2*1);     *200 indicate unequal mean
ability condition;

proc append base=simu.combine&cond1 data=dif4equal; run;
proc append base=simu.combine&cond2 data=dif4unequal; run;

%end;
%end;
%end;
%end;
%mend;
```

A.1.2 Conditions with 40% DIF Items

```
/*----------set parameters----------*/
%let nr=100;
%let ng1=24; %let ng2=40; %let ng3=80;
%let gs1=100; %let gs2=500;

/*---------data generation---------*/
%macro generate8;

%do rep=1 %to &nr;

/*----------parameter matrix----------*/
%let seed11=(800000+&rep*10000+11);
%let seed12=(800000+&rep*10000+12);

proc iml;
  call randseed(&seed11);
  temp_8a0=j(1,20,.);
  call randgen(temp_8a0,'lognormal',0,.3);
  a0=repeat(temp_8a0,8,1);
  use simu.dif8_a; *read in the dif values;
    read all var _all_ into difa;
  close simu.dif8_a;
  a=a0+difa;
  create dif8_a_item1 from a[colname={"a1" "a2" "a3" "a4" "a5" "a6" "a7" "a8"
"a9"
     "a10" "a11" "a12" "a13" "a14" "a15" "a16" "a17" "a18" "a19" "a20"}];
  append from a;
quit;
proc iml;
  call randseed(&seed12);
  temp_8b0=j(1,20,.);
  call randgen(temp_8b0,'normal',0,.5);
  b0=repeat(temp_8b0,8,1);
  use simu.dif8_b;
    read all var _all_ into difb;
  close simu.dif8_b;
```

```
  b=b0+difb;
  create dif8_b_item1 from b[colname={"b1" "b2" "b3" "b4" "b5" "b6" "b7" "b8"
"b9"
     "b10" "b11" "b12" "b13" "b14" "b15" "b16" "b17" "b18" "b19" "b20"}];
  append from b;
quit;

/*----combine a and b parameter, run group replications----*/
%do f1=1 %to 3;       *factor 1, group replication;
  %if &f1=1 %then %let grep=&ng1/8;
  %if &f1=2 %then %let grep=&ng2/8;
  %if &f1=3 %then %let grep=&ng3/8;

%do grouprep=1 %to &grep;

data dif8_combine1;
  merge dif8_a_item1 dif8_b_item1;
  group0=_n_;
  group=group0+(&grouprep-1)*8;
run;

/*------------individual ability------------*/
%do f2=1 %to 2;
  %if &f2=1 %then %let gs=&gs1;
  %if &f2=2 %then %let gs=&gs2;

%let seed13=(800000+&rep*10000+&f1*1000+&f2*100+13);
%let seed14=(800000+&rep*10000+&f1*1000+&f2*100+14);

proc iml;
  call randseed(&seed13);
  theta_equal=j(&gs,8,.);
  beta=j(&gs,8,.);
  theta_unequal=j(&gs,8,.);
  call randgen(theta_equal,'normal',0,1);
  call randgen(beta,'normal',0,1);
  call randgen(theta_unequal,'normal',beta,1);
  create dif8equal_ind1 from theta_equal[colname={"g1" "g2" "g3" "g4" "g5"
"g6" "g7" "g8"}];
  append from theta_equal;
  create dif8unequal_ind1 from theta_unequal[colname={"g1" "g2" "g3" "g4"
"g5" "g6" "g7" "g8"}];
  append from theta_unequal;
quit;
data dif8equal_ind2;
  set dif8equal_ind1;
  person=_n_; meanab=1;
  array g(8) g1-g8;
  do group0=1 to 8;
    theta=g(group0);
      output;
  end;
 run;
data dif8unequal_ind2;
  set dif8unequal_ind1;
  person=_n_; meanab=2;
  array g(8) g1-g8;
```

```
  do group0=1 to 8;
    theta=g(group0);
      output;
  end;
run;
proc sort data=dif8equal_ind2;
  by group0; run;
data dif8equal_combine2;
  merge dif8equal_ind2(drop=g1 g2 g3 g4 g5 g6 g7 g8) dif8_combine1;
  by group0;
run;
proc sort data=dif8unequal_ind2;
  by group0; run;
data dif8unequal_combine2;
  merge dif8unequal_ind2(drop=g1 g2 g3 g4) dif8_combine1;
  by group0;
run;
data dif8_combine3;
  set dif8equal_combine2 dif8unequal_combine2;
run;

/*------------item response------------*/
data dif8_combine4;
  set dif8_combine3;
  call streaminit(&seed14);
    array p(20) p1-p20;
      array r(20) r1-r20;
      array item(20) item1-item20;
      array a(20) a1-a20;
      array b(20) b1-b20;
        do i=1 to 20;
          p(i)=probnorm(a(i)*theta-b(i));
            r(i)=rand('uniform',0,1);
            if p(i)>=r(i) then item(i)=1;
            else item(i)=0;
        end;
run;
data dif8_combine5;
  set dif8_combine4(keep=group person theta a1-a20 b1-b20 p1-p20 item1-item20
meanab);
  rep=&rep;
  dif=8;
run;
proc sort data=dif8_combine5;
  by rep meanab group person;
run;
data dif8_combine6;
  retain dif rep meanab group person;
  set dif8_combine5;
run;

data dif8equal dif8unequal;
  set dif8_combine6;
  if meanab=1 then output dif8equal;
  if meanab=2 then output dif8unequal;
run;
```

```
%let cond1=%eval(8000+100+&f1*10+&f2);
%let cond2=%eval(8000+200+&f1*10+&f2);

proc append base=simu.combine&cond1 data=dif8equal; run;
proc append base=simu.combine&cond2 data=dif8unequal; run;

%end;
%end;
%end;
%end;
%mend;
```

A.2 Data Analysis in Mplus

A.2.1 Example of Random Item Effects Model

```
TITLE:
  random effects

DATA:
  file = data8111.dat;

VARIABLE:
  names = dif rep meanab group person item1-item20;
  usevar = group item1-item20;
  categorical = item1-item20;
  cluster = group;

ANALYSIS:
  type = twolevel random;
  estimator = bayes;
  process = 2;
  biterations = 50000(2000);

MODEL:
  %within%
    s1-s20 | fw by item1-item20;
    fw@1;

  %between%
    item1-item20 s1-s20;
    [s1-s20*1] (p1-p20);
    fb by item1-item20*1(p1-p20);
    fpsi by s1-s20*1(p1-p20);
    fb with fpsi@0;
    fb fpsi;
```

```
OUTPUT:
  !tech1
  !tech8;
  !modindices;

SAVEDATA:
  file = randeffect_fs.dat;
  save = fscores (200);
  !results are randeffect_result.dat;
```

A.2.2 Example of FIXED Alignment Method

```
TITLE:
  FIXED alignment

DATA:
  file = data4111.dat;

VARIABLE:
  names = dif rep meanab group person item1-item20;
  usevar = group item1-item20;
  categorical = item1-item20;
  classes = c(80);
  knownclass = c(group=1-80);

ANALYSIS:
  type = mixture;
  estimator = bayes;
  alignment = fixed(1);
  process = 2;
  biterations = 50000(2000);

MODEL:
  %overall%
    f1 by item1-item20;

PLOT:
  type = plot2;

OUTPUT:
  !tech1 tech8
  align;


!SAVEDATA:
  !file = alignfix_fs.dat;
  !save = fscores (200);
  !results are alignfree_result.dat;
```

A.2.3 Example of FREE Alignment Method

```
TITLE:
  FREE alignment

DATA:
  file = data8111.dat;

VARIABLE:
  names = dif rep meanab group person item1-item20;
  usevar = group item1-item20;
  categorical = item1-item20;
  classes = c(80);
  knownclass = c(group=1-80);

ANALYSIS:
  type = mixture;
  estimator = bayes;
  alignment = free;
  process = 2;
  biterations = 50000(2000);

MODEL:
  %overall%
    f1 by item1-item20;

OUTPUT:
  !tech1 tech8
  align;

SAVEDATA:
  file = alignfree_fs.dat;
  save = fscores (200);
  !results are alignfree_result.dat;
```

A.3 Simulation in SAS

A.3.1 Random Item Effects Model

```
data result1;
  set simu.combine8132;
  ng=80; gs=500;
run;
```

```
%macro random;

%do replication=1 %to 10;

data result1_rep&replication;     *insert &replication;
  set result1;
  if rep=&replication then output;
run;

data result_info(keep=ng gs dif rep meanab);
  set result1_rep&replication;  *insert replication number;
  if _n_=1 then output;
run;

data result2;
  set result1_rep&replication(keep=dif rep meanab group person item1-item20);
run;

data _null_;
  set result2;
  file "&dir\data8132.dat";
  put dif rep meanab group person item1-item20;
run;

x 'Mplus randeffect_fs.inp randeffect_fs.out';

/* read in factor scores and parameter estimates */
data fs1;
  infile "&dir\randeffect_fs.dat";
  input v1-v20 fw_w v22-v25 fw_b v27-v30 fb v32-v35 fpsi v37-v40 bload1 v42-
v45 bload2 v47-v50 bload3 v52-v55 bload4 v57-v60
        bload5 v62-v65 bload6 v67-v70 bload7 v72-v75 bload8 v77-v80 bload9
v82-v85 bload10 v87-v90 bload11 v92-v95 bload12 v97-v100
        bload13 v102-v105 bload14 v107-v110 bload15 v112-v115 bload16 v117-
v120 bload17 v122-v125 bload18 v127-v130 bload19 v132-v135
        bload20 v137-v140 bint1 v142-v145 bint2 v147-v150 bint3 v152-v155
bint4 v157-v160 bint5 v162-v165 bint6 v167-v170 bint7 v172-v175
        bint8 v177-v180 bint9 v182-v185 bint10 v187-v190 bint11 v192-v195
bint12 v197-v200 bint13 v202-v205 bint14 v207-v210
        bint15 v212-v215 bint16 v217-v220 bint17 v222-v225 bint18 v227-v230
bint19 v232-v235 bint20 v237-v240 group;
run;

/* relative bias and RMSD */
data result3;
  merge result1_rep&replication(keep=ng gs dif rep meanab group person theta
a1-a20 b1-b20 p1-p20)
        fs1(keep=fw_w fw_b fb fpsi bload1-bload20 bint1-bint20);
run;

proc append base=simu.V5rand8132_par_rep10 data=result3; run;

%end;
%mend;
```

A.3.2 FIXED Alignment Method

```
data result1;
  set simu.combine8232;
  ng=80; gs=500;
run;

%let gs=500;

%macro align;

%do replication=1 %to 10;

data result1_rep&replication;
  set result1;
  if rep=&replication then output;
run;
data result_info(keep=ng gs dif rep meanab);
  set result1_rep&replication;  *insert replication number;
  if _n_=1 then output;
run;
data result2;
  set result1_rep&replication(keep=dif rep meanab group person item1-item20);
run;
data _null_;
  set result2;
  file "&dir/data8232.dat";
  put dif rep meanab group person item1-item20;
run;

x 'Mplus alignfix.inp alignfix.out';

/* read in parameter estimates */
data load1;    *read in factor loading;
  infile "&dir\alignfix.out" lrecl=100;
  input string $ 1-50;
    if string="MODEL RESULTS" then do;
      do i1=1 to 7;  *skip 7 lines;
          input;
        end;
        do group=1 to 80;   *group number 24,40,80;
          do item=1 to 20;
            input load 23-28;
              output;
          end;
          do i2=1 to 32;  *skip 32 lines;
            input;
          end;
        end;
    end;
run;
data thresh1;     *read in threshold;
  infile "&dir\alignfix.out" lrecl=100;
  input string $ 1-50;
```

```
  if string="MODEL RESULTS" then do;
      do group=1 to 80;   *group number 24,40,80;
        do i2=1 to 32;
          input;
        end;
        do item=1 to 20;
          input thresh 23-28;
            output;
        end;
      end;
  end;
run;
proc transpose data=load1 out=load2;
  by group; id item; var load;
run;
data load3;
  set load2(rename=(_1=load1 _2=load2 _3=load3 _4=load4 _5=load5 _6=load6
_7=load7 _8=load8 _9=load9 _10=load10 _11=load11
       _12=load12 _13=load13 _14=load14 _15=load15 _16=load16 _17=load17
_18=load18 _19=load19 _20=load20));
  drop _name_;
run;
proc append base=simu.V4alignfix8232_load_rep10 data=load3; run; *save loads;
data load4;
  set load3;
  do person=1 to &gs;   *insert &gs here;
      output;
  end;
run;
proc transpose data=thresh1 out=thresh2;
  by group; id item; var thresh;
run;
data thresh3;
  set thresh2(rename=(_1=thresh1 _2=thresh2 _3=thresh3 _4=thresh4 _5=thresh5
_6=thresh6 _7=thresh7 _8=thresh8 _9=thresh9 _10=thresh10
       _11=thresh11 _12=thresh12 _13=thresh13 _14=thresh14 _15=thresh15
_16=thresh16 _17=thresh17 _18=thresh18 _19=thresh19 _20=thresh20));
  drop _name_;
run;
data thresh4;
  set thresh3;
  do person=1 to &gs;   *insert &gs;
    output;
  end;
run;
proc append base=simu.V4alignfix8232_thresh_rep10 data=thresh3; run; *save
thresholds;
/* read in factor scores */
data fs1;
  infile "&dir\alignfix_fs.dat";
  input v1-v20 fmean v22-v225 cg1-cg80 f1_m group;  *need to change with ng;
run;
/* calculate bias */
data result3;
  merge result1_rep&replication(keep=ng gs dif rep meanab group person theta
a1-a20 b1-b20 p1-p20)
        load4 thresh4 fs1(keep=fmean);
```

```
run;
proc append base=simu.V4alignfix8232_combine_rep10 data=results3; run;

/* read in dif detection */
data dif1_a;
  infile "&dir\alignfix.out" lrecl=100;
  input string $ 1-50;
  if string="Class Proportions" then do;
    do i=1 to 87; *24group 31, 40group 47, 80group 87;
        input;
      end;
      do item=1 to 20;
        input name $ 4-15 g1-g80;  *24,40,80;
      output;
    end;
  end;
run;
data dif2_a;
  infile "&dir\alignfix.out" lrecl=100;
  input string $ 1-50;
  if string="Class Proportions" then do;
    do i=1 to 169; *24group 53, 40group 89, 80group 169;
        input;
      end;
      do item=1 to 20;
        input name $ 4-15 g1-g80;   *24,40,80;
      output;
    end;
  end;
run;
/* summerize dif detection results */
data dif1_b;    *intercept;
  set dif1_a;
  array g g1-g80;     *24,40,80;
  array dif_int(80) dif_int1-dif_int80;
  do i=1 to dim(g);
    if g(i)=. then dif_int(i)=1;
      else dif_int(i)=0;
  end;
run;
data dif1_c;
  set dif1_b;
  if sum(of dif_int1-dif_int80)>=1 then dif_int=1;  *24,40,80;
  else dif_int=0;
  keep item dif_int;
run;
proc transpose data=dif1_c out=dif1_d;
  id item; run;
data dif1_e;
  set dif1_d(rename=(_1=dif_int1 _2=dif_int2 _3=dif_int3 _4=dif_int4
_5=dif_int5 _6=dif_int6 _7=dif_int7 _8=dif_int8 _9=dif_int9
      _10=dif_int10 _11=dif_int11 _12=dif_int12 _13=dif_int13 _14=dif_int14
_15=dif_int15 _16=dif_int16 _17=dif_int17 _18=dif_int18
      _19=dif_int19 _20=dif_int20));
  drop _name_;
run;
data dif2_b;     *loading;
```

```
  set dif2_a;
  array g g1-g80;     *24,40,80;
  array dif_load(80) dif_load1-dif_load80;
  do i=1 to dim(g);
    if g(i)=. then dif_load(i)=1;
      else dif_load(i)=0;
  end;
run;
data dif2_c;
  set dif2_b;
  if sum(of dif_load1-dif_load80)>=1 then dif_load=1;    *24,40,80;
  else dif_load=0;
  keep item dif_load;
run;
proc transpose data=dif2_c out=dif2_d;
  id item; run;
data dif2_e;
  set dif2_d(rename=(_1=dif_load1 _2=dif_load2 _3=dif_load3 _4=dif_load4
_5=dif_load5 _6=dif_load6 _7=dif_load7 _8=dif_load8
      _9=dif_load9 _10=dif_load10 _11=dif_load11 _12=dif_load12
_13=dif_load13 _14=dif_load14 _15=dif_load15 _16=dif_load16
      _17=dif_load17 _18=dif_load18 _19=dif_load19 _20=dif_load20));
  drop _name_;
run;
data result9;    *dif result;
  merge result_info dif1_e dif2_e;
run;
proc append base=simu.V4alignfix8232_dif_rep10 data=result9; run;

%end;
%mend;
```

Bibliography

Asparouhov, T., & Muthén, B. (2012). *General random effect latent variable modeling: Random subjects, items, contexts, and parameters.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495-508.

Bauer, D. (2016). A More General Model for Testing Measurement Invariance and Differential Item Functioning. *Psychological methods*.

Bellio, R., & Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling, 5*(3), 217-227.

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

Breslow, N. E., & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika, 82*(1), 81-91.

Byrne, B. M. (1994). Testing for the factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research, 29*(3), 289-311.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological bulletin, 105*(3), 456.

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

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533-559.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4), 355-368.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*: Psychology Press.

Finch, W. H. (2016). Detection of Differential Item Functioning for More Than Two Groups: A Monte Carlo Comparison of Methods. *Applied Measurement in Education, 29*(1), 30-45.

Fox, J.-P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software, 20*(5), 1-16.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*: Springer Science & Business Media.

Fox, J.-P., & Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. *Cross-cultural analysis: Methods and applications*, 467-488.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science, 7*(4), 457-472.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129-145.

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods, 3*(4), 424.

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

Jeffreys, H. (1998). *The theory of probability*: OUP Oxford.

Jöreskog, K. G. (1970). Simultaneous factor analysis in several populations. *ETS Research Report Series, 1970*(2).

Karim, M. R., & Zeger, S. L. (1992). Generalized linear models with random effects; salamander mating revisited. *Biometrics*, 631-644.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American statistical Association, 90*(430), 773-795.

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 524-544.

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, 261-276.

Laird, N. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika, 65*(3), 581-590.

Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 619-678.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*: Routledge.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 86-92. doi:10.1016/0883-0355(89)90024-4

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological methods, 4*(1), 84.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute, 22*(4), 719-748.

McDonald, R. P. (2013). *Test theory: A unified treatment*: Psychology Press.

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

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115-132.

Muthén, B. (2010). Bayesian analysis in Mplus: A brief introduction. *Unpublished manuscript. www. statmodel. com/download/IntroBayesVersion, 203*.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological methods, 17*(3), 313.

Muthén, B., & Asparouhov, T. (2013a). BSEM measurement invariance analysis. *Mplus Web Notes, 17*, 1-48.

Muthén, B., & Asparouhov, T. (2013b). New methods for the study of measurement invariance with many groups. *Mplus. statmodel. com [12.04. 2014]*.

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: the alignment method. *Frontiers in psychology, 5*, 978.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*(3), 376-398.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Vol. 161): Sage Publications.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education, 14*(3), 235-259.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495-502.

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

Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. *Copenhagen: Danish Institute for Educational Research*.

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

Roju, N. S., Van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied psychological measurement, 19*(4), 353-368.

Rudner, L. M., & Gagne, P. (2001). *An overview of three approaches to scoring written essays by computer*: ERIC Clearinghouse on Assessment and Evaluation.

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*(6), 1292.

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross‐national consumer research. *Journal of consumer research, 25*(1), 78-107.

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

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393-408.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological bulletin, 99*(1), 118.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American statistical Association, 81*(393), 82-86.

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486-492.

van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*: Springer Science & Business Media.

Verhagen, A., & Fox, J. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology, 66*(3), 383-401.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*(3), 532-547.