DETECTION OF LATENT DIFFERENTIAL ITEM FUNCTIOING (DIF) USING MIXTURE 2PL IRT MODEL WITH COVARIATE

by

Ya Zhang

B.A., East China Normal University, 2004

M.A., East China Normal University, 2007

M.A., University of Pittsburgh, 2015

Submitted to the Graduate Faculty of

the School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Ya Zhang

It was defended on

August 16, 2017

and approved by

Suzanne Lane, Professor, Department of Psychology in Education

Clement Stone, Professor, Department of Psychology in Education

Feifei Ye, Assistant Professor, Department of Psychology in Education

Lauren Therhost, Associate Professor, Department of Occupational Therapy

Dissertation Advisor: Suzanne Lane, Professor, Department of Psychology in Education

Clement Stone, Professor, Department of Psychology in Education

DETECTION OF LATENT DIFFERENTIAL ITEM FUNCTIOING (DIF) USING MIXTURE 2PL IRT MODEL WITH COVARIATE

Ya Zhang, PhD

University of Pittsburgh, 2017

Mixture IRT models have been shown to improve the identification of latent group structure and facilitate the estimation of model parameters when covariates are incorporated or the Bayesian estimation method is employed. However, the efficiency of mixture IRT models in DIF analysis has not been systematically studied due to the challenges of identifying DIF with a relatively complex model. The present dissertation aims to explore the effect of covariate and estimation method on the detection of latent DIF under the mixture IRT framework. A Monte Carlo simulation study was performed by manipulating the magnitude of DIF, type of DIF, proportion of DIF items, group impact, and relationship between the covariate and the latent group membership. The generated response data were analyzed using the mixture 2PL IRT model by manipulating the inclusion of covariate and the estimation method. The estimation results were evaluated in terms of the recovery of the latent group structure, recovery of the model parameters, and detection of DIF. The goal is to provide insights and suggestions on the use of mixture IRT models in the analysis of DIF.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

The basis of this research stemmed from my interest in exploring statistical models for measurement invariance. Not only used as a measure of students' academic achievement, test scores are often relied on in making high-stakes decisions. However, without the support of measurement invariance equality or fairness is hard to achieve in the selection of individuals. It became interesting to me to find out what statistical models would work effectively in identifying items that hamper the fairness of tests.

This dissertation represents the combined efforts of a strong support group. I could not find a single word to convey my gratefulness to my advisor, Dr. Suzanne Lane. I am heartily thankful to her not only because she is a role model as a scholar and has been giving me opportunities of exploring topics of my interest, but also because she has been giving me encouragement, caring, and guidance in all the fields throughout my doctoral study. She proof-read multiple versions of this dissertation and provided many substantive suggestions to help improve the clarity of my arguments. Without her valuable advice, I would not be able to complete this dissertation research in a timely manner. She inspires me to pursue my career goal in academia, and I could not find a mentor better than her. My deepest appreciation goes to my co-advisor Dr. Clement Stone. He taught me the knowledge about simulation study and the Bayesian theory in a very understandable way. He gave me many suggestions to improve the study design of this work, and was patient to solve all my questions about coding and technical issues. I learnt from him to think like a researcher and always keep an inquisitive and creative mind.

I also would like to express my sincere gratitude to Dr. Feifei Ye. She is a teacher and more of a friend of mine. She is a steady and generous source of guidance, feedback, and help, and she has been very supportive throughout my time at Pitt. She is always willing to share her thoughts and give me suggestions on my dissertation, and always made herself available to answer my questions. The care and friendship from her make my doctoral study a very pleasant and unforgettable memory. I am thankful to Dr. Lauren Therhost for serving as my committee member and giving me suggestions on writing my dissertation.

Finally, I am grateful for all the other faculty members, my graduate colleagues in the research methodology program, and my beloved family. I feel no fear of failure and never feel alone on my way to becoming a scholar all because of your constant encouragement and unwavering support.

1.0 INTRODUCTION

The assessment of the presence of differential item functioning (DIF) is a key component for test development and validation. An item is said to exhibit DIF when the item functions differently in a focal group in comparison with a reference group after controlling for differences in levels of performance on a latent trait (e.g., ability) of interest (Holland & Wainer, 1993; Scheuneman, 1979). In psychometric practice, measurement experts usually investigate differential item functioning (DIF) for demographic groups to help ensure that tests are fair. The presence of DIF is considered a serious threat to test validity because it implies that one group has an unfair advantage on an item in comparison with another group. Therefore, it often leads to the inequity of comparing conditions, invalid comparison of group differences, and unfairness of selection (Meredith, 1993; Millsap & Kwok, 2004).

Since DIF may result in serious consequences, it has generated extensive research from different perspectives and many researchers focused on the methods of DIF identification. The traditional view is that DIF is detected based on examinees' group membership that can be easily observed such as gender and ethnicity. However, it is known that group membership is not always defined in terms of evident/observable attributes. Examinees can form as a homogeneous group based on certain unobservable features. Recently, a surge of research has emerged to investigate DIF across latent (unknown) groups as opposed to manifest groups, and to this end mixture IRT modeling has gained attention as a promising tool. The current simulation study was built upon this work, and this study will contribute to the literature regarding the detection of latent DIF based on mixture IRT models.

1.1 STATEMENT OF PROBLEM

The arguments for the inappropriateness of using the conventional approach to study DIF were brought up by several studies on the analysis of DIF using mixture IRT models (Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton, 2002). The main issue is that examinees for whom items functioned differently cannot be accurately characterized by their manifest grouping variable, despite that the groups are defined in terms of such a grouping variable. Furthermore, an item set that does not exhibit DIF with respect to the manifest group variable does not guarantee zero DIF within this item set. This is because the detection of DIF based on manifest dimensions fails to account for secondary dimensions, also known as nuisance dimensions (Bolt & Stout, 1996; Oshima & Miller, 1992; Roussos & Stout, 1996). Thus, the consideration for latent DIF is important given that the manifest group membership may not be reliable or provide a valid indication of examinees' true group membership.

Mixture IRT models have been widely used to assess DIF, which allows for a comparison of item parameters across latent groups. This idea is similar to multigroup IRT models, except that the groups are recovered as a part of model estimation rather than specified a priori. Empirical and simulation evidence support the use of mixture IRT models in DIF detection (Cho, Suh, & Lee, 2015; Cohen & Bolt, 2005; De Ayala et al., 2002; Maij-de Meij, Kelderman, & van der Flier, 2010; Samuelsen, 2005). Based on Smit, Kelderman, and van der Flier's discussions on the positive effect of collateral information on item parameter estimation and latent class membership assignment (1999, 2000), researchers have begun to investigate the methodological impact of covariates on IRT mixture modeling (Chen & Jiao, 2014; Choi, Alexeev, & Cohen, 2014; Dai, 2009; Li, Jiao, & Macready, 2015; Tay, Newman, & Vermunt, 2011). Theoretically, covariates can be involved as predictors at each latent class level or as predictors of the latent class membership. Most of these studies suggest that the inclusion of covariates has the potential to improve latent class identification and facilitate the interpretation of differences across latent groups.

Simulation studies that used latent DIF detection based on mixture IRT models identified a set of conditions that may affect the behavior of mixture IRT models in detecting DIF. These simulation conditions include sample size, test length, number of latent classes, the overlap between covariates and latent classes, mean difference in ability between latent groups (impact), percentage of DIF items, and magnitude/effect size of DIF.

Unfortunately, none of these studies provided a thorough discussion on the IRT mixture modeling of latent DIF. Several unresolved issues include: 1) More complex mixture IRT models were infrequently studied. The mixture Rasch model, as the simplest mixture IRT model, was adopted in most of these studies. 2) Non-uniform DIF across latent groups was not addressed. In the case of non-uniform DIF, members of one group are favored up to a level on the ability scale and from that point on the relationship is reversed. Although non-uniform DIF is not as common as uniform DIF in practice, it still occurs in testing programs. 3) Effect of estimation methods on latent DIF detection was not considered. Finch and French (2012) compared the effects of MLE and MCMC on the classification accuracy and parameter estimation of the mixture 1PL and 2PL IRT models. It was found that MCMC estimation produced better recovery of the latent group membership and more accurate parameter estimates for smaller samples and fewer items. In contract, MLE provided more accurate parameter estimates for larger samples and more items. A direct comparison of the two procedures is needed to provide insights into the optimal estimation conditions in latent DIF detection.

1.2 PURPOSE OF THE STUDY

The main purpose of this dissertation was to conduct a simulation study to address the performance of mixture IRT models in assessing DIF across latent groups by highlighting the role of estimation methods in detecting two different types of DIF. Mixture 2PL IRT models were used to analyze the generated datasets, which offered the flexibility of analyzing latent DIF in terms of item difficulty and item discrimination. As a dominant estimation procedure in previous simulation work, the Bayesian MCMC has demonstrated satisfactory item parameter estimation and classification accuracy. On the other hand, DeMars and Lau (2011) found that the recovery of class membership was poor across all the simulation conditions using FIML estimation. The time-consuming issue of the MCMC estimation was non-trivial for mixture IRT models. The difficulty of convergence to solutions for individual parameters was also of concern (Li, Cohen, Kim, & Cho, 2009). Compared to MCMC estimation, MLE is generally more time efficient but could still have other problems such as the local maxima issue. In that case, longer computation time is needed due to the maximum number of iterations needed (Finch & French, 2012). Therefore, it is not clear which estimation gives an optimal condition in latent DIF detection without a direct comparison of the two methods. The main purpose of the present study was to address the differences of MLE and MCMC estimation in the mixture IRT modeling of latent DIF.

Seven factors were manipulated in the present study which were DIF size, the percentage of DIF items, DIF type, group impact, the association of the covariate with the latent group membership, analyzing model, and estimation method. The first five factors were the between-replication factors, and the last two were the within-replication factors. All the factors were dichotomous to decrease the complexity of the design. Specifically, this study aims to answer the following research questions: 1) How do the simulation factors (DIF size, the percentage of DIF items, DIF type, group impact, the association of the covariate with the latent group membership, model, and estimation method) affect the performance of the mixture 2PL model in latent DIF detection? 2) How is the mixture modeling of latent DIF affected by the estimation method employed (MLE v.s. Bayesian estimation)? 3) How well are the latent DIF items detected under disadvantaged conditions (i.e., small number of DIF items, small DIF size, and weak association between the covariate and the latent group membership)?

1.3 OVERVIEW OF CHAPTERS

In Chapter 2, the theoretic frameworks of the mixture IRT model and mixture IRT model with covariate are described, followed by an introduction to the estimation methods used in the mixture IRT literature. The most frequently used algorithms for each estimation method are summarized. The estimation difficulties associated with each estimation procedure is discussed along with a summary of the solutions to these challenges. The last part of this chapter summarizes previous research on latent DIF detection with a focus on the findings from simulation studies.

Chapter 3 presents the design of this simulation study. The selection of each of the stimulation factors is justified using relevant simulation work. The measures of outcomes are described, followed by a discussion about the solutions to the label switching issue. Each of the simulation steps are described. The validation of simulation parameters is also described, and the validation results are presented in Appendix A.

The results of the simulation study are presented in Chapter 4, which includes the description of the analytical plan, the discussion of the convergence and label switching issue, as well as a summary of the results for each of the outcome measures.

Chapter 5 summarizes the findings, and discusses the potential limitations and future directions for the use of mixture IRT models in latent DIF analysis.

2.0 LITERATURE REVIEW

Mixture IRT models with covariates (Mixture IRT-C) proposed in this study are used for the detection of DIF by employing two different estimation methods. This chapter begins with a discussion of mixture IRT models, from the simplest to the full parameter format. The next section explains mixture IRT-C models with an emphasis on the benefits for parameter estimation and latent class assignment based on theoretical and empirical evidence. Model parameter estimation for mixture IRT models have been explored by using both MLE and MCMC procedures. The third section compares the two techniques by summarizing studies using each estimation method within the mixture IRT framework. The following two sections identifies the major difficulties involved in latent modeling. The detection of latent DIF based on mixture IRT-C models are evaluated. The last section describes the manifestation of DIF between latent groups and reviews both the empirical and simulation studies on latent DIF analysis using mixture IRT models.

2.1 MIXTURE IRT MODELS

Item response theory (IRT) models contain a family of models that describe the probability of the response behavior of an examinee given his/her personal characteristics (i.e., mathematical ability) and the item difficulty. Two essential assumptions that IRT models require are unidimensionality and local independence (Hambleton & Swaminathan, 1985; Lord, 1980; Lord & Novick, 1968; Reckase, 1979). Specifically, the latent construct measured in a test presents in one-dimensional space only, and individuals’ response to test items are independent of one another such that the probability of answering an item correctly only depends on the latent trait. In the situation where individuals' cognitive strategies are different, the conventional IRT assumptions would be violated by introducing secondary dimension(s), known as population heterogeneity. Mixture IRT models are developed to take qualitative individual differences into account without losing necessary assumptions of the basic models (Rost, 1990).

Theoretically speaking, mixture IRT models are considered as the integration of finite mixture models and conventional IRT models. The early work can be traced back to the late 1980s when research mainly focused on integrating Rasch models and latent class analysis (LCA) (Clogg, 1988; Formann, 1985, 1989). The basic idea is that Rasch models describe the response behavior within a latent class, and different parameter values are allowed to be estimated for different latent classes of examinees in a population (Mislevy & Verhelst, 1990; Rost, 1990; von Davier & Rost, 2006). In particular, mixture IRT models identify latent memberships and establish item response functions within each class simultaneously (De Ayala & Santiago, 2016). An examinee population is considered to be composed of a fixed number of latent groups (Cohen, Wollack, Bolt, & Mroch, 2002). Homogeneity is assumed within each latent group such that examinees of the same latent group share unique characteristics and model parameters can differ across latent groups. The assumption of a single qualitatively homogeneous distribution of IRT models is relaxed, and the assumption of same response probability is also relaxed (Cho, 2013).

Mixture IRT models take the weighted sum of the probability of the correct response across the latent groups, which can be represented as:

$$P\left(y_{ij} = 1 \middle| \theta_i\right) = \sum_{g=1}^{G} \pi_g \, P\left(y_{ijg} = 1 \middle| \theta_i, g\right), \qquad (1)$$

where π_g is the proportion of examinees belonging to the latent group g and is constrained such that $\sum_{g=1}^{G} \pi_g = 1$; $P(y_{ijg} = 1|\theta_i, g)$ is the probability of person *i* correctly responding to item *j* given his/her latent group membership *g* and ability level θ_i.

The simplest mixture IRT model is the mixture Rasch model (MRM), which was extensively used in the literature. The early work that contributed to the development of MRM were conducted by Kelderman and Macready (1990), Mislevy and Verhelst (1990), and Rost (1990). In these studies, MRM was derived based on the same idea of combining a latent trait with a latent categorical variable. Parameters to be estimated in this model are Rasch difficulty and class-specific ability parameters:

$$P(y_{ij} = 1|\theta_i) = \sum_{g=1}^{G} \pi_g \frac{\exp(\theta_{ig} - b_{jg})}{1 + \exp(\theta_{ig} - b_{jg})}, \quad (2)$$

where θ_{ig} is the ability level of person *i* within the latent group *g*; b_{jg} is the difficulty of item *j* for the latent group g. It is noted that in the literature the subscript of the latent group membership g is often omitted for θ_{ig} because individuals have only one ability parameter given their latent group membership.

By adding one more parameter into the model, the mixture 2-parameter logistic (2PL) IRT model relax the discrimination assumption and can be viewed as an extension of MRM. For each of the latent groups a 2PL model is assumed to hold while the item difficulty and discrimination parameters are allowed to be different. Existing research investigated the accuracy of identifying the correct number of latent classes by comparing across different mixture IRT models (Li, Cohen, Kim, & Cho, 2009; Sen, Cohen, & Kim, 2014). The recovery of latent membership was found to be the best for the mixture 2PL model regardless of the test length. Particularly, as a more complex model the mixture 2PL model tended to yield fewer spurious latent class solutions due to latent nonnormality. The probability of a correct response in mixture 2PL model takes the same form as the MRM and can be represented as:

$$P(y_{ij} = 1|\theta_i) = \sum_{g=1}^{G} \pi_g \frac{\exp[a_{jg}(\theta_{ig} - b_{jg})]}{1 + \exp[a_{jg}(\theta_{ig} - b_{jg})]}, \tag{3}$$

where a_{jg} is the discrimination of item *j* for the latent group g.

Similarly, the mixture 3-parameter logistic (3PL) IRT model assumes that the 3PL model holds for each of the latent groups with the item difficulty, discrimination, and guessing parameters permitted to be different across the latent groups. The mixture 3PL model was applied only in a few studies (Cohen & Bolt, 2005; Li et al, 2009; Sen, Cohen, & Kim, 2014). The probability of a correct response in mixture 3PL model is represented as:

$$P(y_{ij} = 1|\theta_i) = \sum_{g=1}^{G} \pi_g \, [c_{jg} + (1 - c_{jg}) \frac{\exp[a_{jg}(\theta_{ig} - b_{jg})]}{1+\exp[a_{jg}(\theta_{ig} - b_{jg})]}], \tag{4}$$

where c_{jg} is the lower asymptote parameter of item *j* for the latent group g.

An important feature of mixture IRT models is that the heterogeneity of response data among the latent classes and within each latent class can be well represented. The class membership is considered to reflect the qualitative differences of the response patterns across latent groups, while the ability is considered to reflect the quantitative differences among individuals within a latent group. Mixture models provide more flexibility but they are criticized for some limitations. For example, the number of parameters estimated increases rapidly as the number of latent classes increase. This is because both the ability and membership parameters for each examinee as well as the item parameters for each class need to be estimated (Choi, 2010). This would make the model difficult to interpret when more than two latent classes are identified. In that case, a large sample size is needed to cope with the increased number of parameters to be estimated (Li, Cohen, Bottge, & Templin, 2015). In addition, mixture IRT models by themselves do not explain the latent categorizing of examinees. Therefore, collateral/additional information is recommended to be included to improve the understanding of how and what causes the latent groups to be different (Smit et al., 1999).

2.2 MIXTURE IRT MODELS WITH COVARIATES

As was stated above, one of the most challenging tasks of using mixture IRT models is to determine what causes the heterogeneity of the population, which is also a question researchers often ask when using latent class modeling. It has been widely shown that collateral information such as background characteristics may confound the estimation of the modeled variables. The inclusion of these potentially effective covariates may help alleviate the difficulties of identifying latent class and improve the estimating of model parameters. Evidence about the benefits of covariate inclusion were found in both the mixture modeling and conventional IRT literature. This evidence is discussed in this section, followed by a description of the methods of including covariates.

2.2.1 Covariate Effect in Mixture Models

In general, there are two ways that covariates are included in a latent variable mixture model. They can be associated with the underlying class variable or the measured outcome variable. Several early studies modeled and analyzed the relationship between covariates and latent class membership. For example, Dayton and Macready (1988) developed a concomitant-variable latent class model where covariates were used to predict the class membership. Van der Heijden, Dessens, and Bockenholt (1996) further discussed the concomitant-variable latent class model where continuous explanatory variables were added and maximum likelihood estimates of the parameters through the EM algorithm were derived.

Recent studies discussed the effect of including covariates into LCA and growth mixture modeling. To examine the covariate effect on parameter coverage and class membership assignment, Lubke and Muthén (2007) conducted a simulation study where the class separation, measured by the multivariate Mahalanobis distance, and the effect size of a single continuous covariate were manipulated. It was found that parameter coverage was acceptable even for the smallest class separation, and the class assignment seemed to be improved when increasing the class separation or covariate effect size. In other words, the inclusion of covariates into the LCA model contributed to the reduction of classification error rates. As an extension of these findings, Wurpts and Geiser (2014) examined the ways that covariates influence the performance of LCA. Five factors were manipulated in this simulation study including the sample size, number of latent classes, number and quality of latent class indicators, as well as the effect size of a continuous covariate on latent class membership. It was found that the large covariate effect size improved the prediction of latent class membership. More interestingly, the large covariate effect size somehow compensated for the small number and the low quality of class indicators. These studies were consistent with previous research in that adding covariates is beneficial to LCA estimation process and the additional information provided by covariates could offset some suboptimal model conditions.

Covariate effect was also addressed within the growth mixture modeling (GMM) framework. Muthén (2004) found that covariates improved model specification and class membership assignment when used as a predictor of trajectory class membership in GMM. In particular, class membership assignment was found to be influenced by the association between the covariate and the trajectories as well as the distribution of the covariate. For example, Huang, Brecht, Hara, and Hser (2010) examined the agreement on the latent trajectory class membership classification between the unconditional GMM model (the model without the covariate) and the conditional GMM models (the models with the covariates such as the early heroin use). When the distribution of the covariate was highly unbalanced, meaning a majority of subjects were in one group (i.e., 81.5% of the subjects were the non-early-heroin users and 18.5% were the early-heroin users), the inclusion of the covariate did not change subjects' trajectory group membership substantively. This was because it contributed little extra information to the classification of subjects into different trajectory groups (i.e., only a small number of subjects were affected by the inclusion of early heroin use). In contrast, subjects' class membership changed substantively when the covariate had relatively low correlation with the latent trajectory membership and had a balanced distribution. GMM with correctly specified covariates outperformed the model without covariates in recovering the correct number of classes in terms of various model fit indices. However, the misspecification of covariates led to the deviation of GMM from correct number of classes (Li & Hser, 2011). In addition, the inclusion of covariates was recommended to improve class membership assignment when the study sample size and the class separation were small (Hu, Leite, & Gao, 2017). This was consistent with previous findings in that the inclusion of covariates compensated for the inadequacy of GMM.

2.2.2 Covariate Effect in IRT Models

Covariate effect in the IRT framework generally follows the same pattern as the one discovered in mixture models in that the precision of item parameter estimation is improved by including auxiliary information about examinees (i.e., age, education, demographic information, etc.).

Specifically, collateral information about examinees contributes to reducing the standard errors in the estimation of item parameters and the mean squared errors in the estimation of individual proficiencies/abilities (Mislevy, 1987; Mislevy & Sheehan, 1989). There was also evidence showing that the inclusion of a collateral variable that was highly related to the ability parameter contributed substantially to the reduction of the mean squared error in ability prediction but was less useful in improving the estimation of item parameters (Adams, Wilson, & Wu, 1997). In order to produce consistent item parameters, the collateral information used in item parameter estimation needed to be used in item selection as well (Mislevy & Sheehan, 1989).

In the literature, IRT models with covariate (IRT-C) was mainly used to assess DIF. In a simulation study by Tay, Vermunt, and Wang (2013) a single covariate was included into the 2PL IRT model, which was compared with two other procedures in identifying DIF including the Mantel-Haenszel and the multiple indicators multiple cause procedures. The authors examined the power and Type I error rates for DIF detection, the performance of the IRT-C model in ascertaining different types of DIF (uniform and nonuniform), as well as the deviance of the focal group latent mean from the simulated value. It was found that the IRT-C model performed the best with the highest power and well-controlled Type I error rate at .05 when there was a moderate number of DIF items in the test. In this condition, the IRT-C model produced accurate estimates of the focal latent mean in that the root mean squared error (RMSE) was as low as 0.10 across all the simulated conditions. A follow-up study by Tay, Huang, and Vermunt (2016) examined the power of detecting DIF and the recovery of latent means when multiple covariates were included in the 3PL IRT model. This study basically verified the previous finding that the model parameters (particularly the item discrimination and item difficulty parameters) and the latent means were accurately estimated with a substantial sample size of 20, 000 regardless of the DIF effect.

2.2.3 Inclusion of Covariates in Mixture IRT Models

Most covariates considered in the mixture IRT models are manifest variable, which are typically used as the predictors of latent class membership π_g through a logistic regression function or the predictors of individual latent trait θ_{ig} through a linear regression function (Dai, 2013; Li et al., 2015; Tay et al., 2011):

$$logit(\pi_g) = \beta_{0g} + \beta_{1g}c_i, \tag{5}$$

or equivalently,

$$\pi_g = \frac{\exp(\beta_{0g}+\beta_{1g}c_i)}{\sum_{g=1}^{G}\exp(\beta_{0g}+\beta_{1g}c_i)}, \tag{6}$$

and,

$$\theta_{ig} = \alpha_{0g} + \alpha_{1g}c_i + e_{ig}, \tag{7}$$

where c_i denotes the covariate for examinee *i*; β_{0g} and β_{1g} are the regression coefficients in the logistic regression; *G* denotes the number of latent classes (i.e., for two latent classes *G* = 2); α_{0g} and α_{1g} are the intercept and slope of the latent regression model for the latent group *g*; e_{ig} is the error term of the linear function with a distribution of *N*(0, σ_{eg}^2).

The graphical representation of the relationship among latent group, ability, and covariate was proposed by Tay et al (2011). As shown in Figure 2.1, g is the latent class, θ is the latent trait, y_j is the observed indicator, and c is the vector of observed characteristics of examinees. Path 1 and 2 represent the prediction of the latent class membership and the latent trait by the covariate respectively. Path 3 and 4 represent the dependence of examinees' responses on their latent group membership and observed characteristics, which correspond to the latent DIF and observed DIF respectively. Path 5 denotes the dependence of examinees' latent trait on their latent class membership, which is usually defined by the mean difference of abilities between the latent groups.

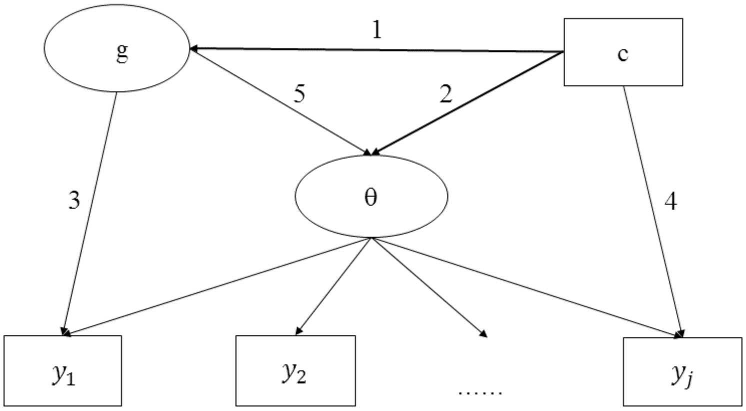


Figure 2.1. Graphic Representation of Mixture IRT Model with Covariate

Samuelsen (2005) summarized three types of variables that may be of interest and informative as covariates in educational research: 1) non-traditional manifest grouping variables such as native speakers versus non-native speakers, urban versus non-urban students, etc.; 2) continuous predictors such as the number of math classes a student has taken and the number of years an English language learner has been in the United States; 3) interactions between traditional and non-traditional manifest grouping variables such as the interaction between race or gender and geographical location.

The inclusion of covariates are expected to achieve the same goal in mixture IRT models as they are in LCA and conventional IRT models. Several studies have discussed the performance of mixture IRT-C models and generally verified the positive role of covariates in the recovery of underlying structures and parameter estimation within that structure. Specifically, the percentage of correct classification was higher for the model with covariates correctly specified, which also produced the smallest SE and RMSE in item parameter recovery. Smit, Kelderman, and van der Flier (1999, 2000) evaluated the usefulness of incorporating covariates into mixture Rasch models. In their studies, the strength of the association between the covariates and the latent class variable was manipulated. It was found that the mean difference between the simulated parameters and the estimated parameters, known as the deviance, became smaller as the external variables became more associated with the latent class membership. The standard deviation of the deviance reduced as more collateral information was included in the model. The percentage of correct assignment of subjects also became higher as the association between the latent class and the covariate became stronger, especially for relatively large sample sizes.

Samuelsen (2005) judged the appropriateness of detecting DIF using a manifest group approach when the manifest group membership did not completely overlap with the latent group membership. In this study, different background variables were incorporated into a mixture Rasch model as the indicators of the latent class membership for the manifest groups. Better item parameter recovery was found when the covariate was strongly associated with the latent class, especially when the sample size was small. Other studies also demonstrated that the inclusion of covariates improved the recovery of item difficulty and group membership and facilitated the discovery of underlying structures when sample sizes or the difference between latent classes were small (Cho, Cohen, & Kim, 2006; Cho, Cohen, Kim, & Bottge, 2010; Li, Cohen, & Bottge, 2007).

Most recently, Li, Jiao, and Macready (2015) studied different approaches of incorporating covariates into mixture IRT models. Dichotomous variables were used as the predictors of the probability of an examinee belonging to a latent class, while continuous variables entered the model as the predictors of latent ability. It was found that the item parameter recovery was generally better for the mixture Rasch model with covariates, especially for the model with the continuous covariate only. In addition, the mixture Rasch model with continuous covariate resulted in smaller standard errors of the person parameter estimates and the best person parameter recovery, which had negligible difference from the true model. The model with both the dichotomous and continuous covariates led to the most accurate latent class assignment. The model with one dichotomous covariate performed slightly better than the model with one continuous covariate. Lastly, the study suggeste the deviance information criterion (DIC) to be the most effective model fit index, which selected the mixture Rasch model with correctly specified covariate, whereas other fit indices such as Bayesian information criterion (BIC), consistent Akaike's information criterion (CAIC), or sample-size adjusted BIC (SABIC) favored the parsimonious rule and preferred the model without covariates.

2.3 ESTIMATION OF MIXTURE IRT MODELS

Maximum likelihood and Bayesian analysis with Markov Chain Monte Carlo (MCMC) sampling are the most frequently used estimation methods for mixture IRT models. A review of the literature found that different types of maximum likelihood were involved, including conditional maximum likelihood (CML) (Rost, 1990, 1991), joint maximum likelihood (JML) (Willse, 2010), and marginal maximum likelihood (MML) (Mislevy & Wilson, 1996; Von Davier & Yamamoto, 2004). As an alternative technique, the MCMC estimation involves an iteration process based on the model's parameter space, which was employed in many studies on mixture modeling. The derivation of likelihood functions are not required in MCMC estimation, whereas the specification of parameter prior distributions is an essential part of the MCMC algorithm. In this section the algorithms/schemes of the maximum likelihood and Bayesian estimation are described with a focus on the MCMC estimation, followed by a discussion on their applications in the mixture IRT research.

2.3.1 Maximum Likelihood Estimation

All three maximum likelihood methods have been used in IRT parameter estimation, and they differentiated in terms of the likelihood being maximized. Compared to JML, MML and CML are more commonly used for mixture IRT models (Mislevy & Wilson, 1996; Rost, 1990, 1991; Rost & von Davier, 1995; von Davier & Rost, 1995; von Davier & Yamamoto, 2004). The JML procedure treats both the item and the ability as unknown but fixed model parameters. It finds the model parameters by jointly maximizing the following likelihood across all the examinees and all the items, where ϕ and θ are the item and ability parameters respectively:

$$L(\phi, \theta | X) = \Pi L(\theta | x, \phi). \tag{8}$$

Unlike JML, MML assumes that item parameters are fixed but examinees' abilities are randomly sampled from some larger distribution. It integrates over the distribution of the ability parameter in the likelihood equation. MML finds the item parameters by maximizing the following likelihood without reference to the ability parameter, where $\Pr\{x|\phi\}$ is the marginal/unconditional probability of observing the item response vector and $g(\theta)$ is a continuous ability distribution (Johnson, 2007):

$$L(\phi | X) = \prod \Pr\{x | \phi\} = \int_{\Theta} L(\theta | x, \phi) g(\theta) d\theta. \tag{9}$$

Rather than making an assumption about the distribution of the latent variable, another solution to ensure consistent item parameter estimation is to introduce the sum of the correct responses/raw score for each examinee as a sufficient statistic for an individual's ability. The item parameters can be estimated through the conditional distribution of the responses given the scores, and therefore, the probability of the responses no longer depends on the value of θ. CML finds the item parameters by maximizing the following conditional likelihood, where T is the vector of the observed raw scores (Eggen, 2000; Johnson, 2007):

$$L(\phi|X,T)=\prod \Pr\{x|t,\phi\}. \tag{10}$$

2.3.1.1 Expectation–Maximization (EM) Algorithm

Given the observed response pattern X, the latent attribute θ, and a vector of unknow parameters ϕ, along with a likelihood function, the expectation–maximization (EM) algorithm was wildly used in mixture IRT literature. There are minor algorithm differences in the estimation of class-specific model parameters, depending on the implementation of maximum likelihood (JML, MML, or CML). In general, the EM algorithm is an iterative procedure including two steps in each iteration. In the expectation step or the E-step, the task is to calculate the expected value of the log-likelihood of the parameter given the observed response pattern. In the maximization step or the M-step, the task is to choose the parameters that maximize the expected log-likelihood obtained from the E-step. This process continues for the E- and M-steps until the resulting maxima for the parameters change very little compared to the ones from the previous M-step.

Using the mixture Rasch model as an example, Rost (1990, 1991) summarized the use of EM algorithm in mixture IRT models. First, the E-step calculates the expected frequencies of the response pattern for each latent class by weighting the observed response frequencies with the probability of the response belonging to that class, also known as the within-class frequencies (Willse, 2010):

$$\hat{n}(x|c)=n(x)\frac{\pi_c P(x|c,\theta_c,b)}{P(x)}=n(x)\frac{\pi_c P(x|c,\theta_c,b)}{\sum_{c=1}^{C}\pi_c P(x|c,\theta_c,b)}, \tag{11}$$

where *x* is the response vector, *c* denotes the latent class, π_c is the probability of the latent class *c*, θ_c is the ability for the latent class, and *b* is the item parameter. $P(x|c,\theta_c,b)$ is the probability of examinees' responses to test items given their latent class membership, abilities corresponding to that latent class, and the item difficulty. $P(x)$ is the probability of the response vector across the latent class membership.

Second, based on the within-class frequencies from the E-step, the M-step estimates the item and ability parameters for each latent class by maximizing the log-likelihood function of the observations in that class. Item parameters are computed by setting the first partial derivatives of this function to be zero. The probability of the latent class membership is simply calculated by the following equation, where N is the sample size:

$$\hat{\pi}_c = \frac{\sum_x \hat{n}(X|C)}{N}. \tag{12}$$

2.3.1.2 Use of MLE in Mixture IRT Models

In addition to the theoretical discussion of the use of MLE in mixture IRT models (Mislevy & Wilson, 1996; Rost, 1990, 1991; Willse, 2010), a number of empirical studies employed the EM algorithm within the mixture IRT framework, and most of these studies were on DIF analysis (Aryadoust, 2015; Chen & Jiao, 2014; Cho, 2013; DeMars & Lau, 2011; Kelderman & Macready, 1990; Maij-de Meij, Kelderman, & van der Flier, 2010; Tay, Newman, & Vermunt, 2011; Van Nijlen & Janssen, 2008). A few others were on the usefulness of collateral information incorporated in mixture IRT models (Smit, Kelderman, & van der Flier, 1999, 2000) and the comparability of the profiles across latent groups (Aryadoust & Zhang, 2016; Paek & Cho, 2015).

The first study that explained DIF based on mixture IRT models was done by Kelderman and Macready (1990), where a loglinear format of the Rasch model, termed loglinear latent class model, was used to detect the presence of both the manifest and latent DIF. MLE with the iterative proportional fitting (IPT) algorithm was used in this study to estimate model parameters. A practical problem discovered in this study was that a large number of iterations might be needed to reach a solution, especially when a complex model was used or the initial values of the interactive process were not reasonable. Adopting this general framework, Smit, Kelderman, and van der Flier (1999, 2000) conducted two simulation studies to examine the impact of collateral information on item parameter estimation in mixture Rasch model and Birnbaum's two-parameter model. The E-M algorithm was used in both of the studies with the E-step computing the expected frequencies given the observed data and current parameter estimates and the M-step maximizing the log-likelihood.

In a more recent study mixture IRT models with covariates were used to identify overall DIF (Tay et al., 2011). The authors recommended the use of multiple random start values to avoid the multiple local maxima issue for the log likelihood and the increasing of quadrature points to improve the estimation accuracy. In addition, full information maximum likelihood (FIML) was used in some studies with missing data (Chen & Jiao, 2014).

Lastly, through imposing the item difficulty equality constraint and ability distribution constraint using different programs (WINMIRA v.s. Mplus), Paek and Cho (2015) found that the CMLE employed in WINMIRA did not produce noticeable estimation differences compared to the MMLE employed in Mplus. However, the choice of estimation algorithm depends on the program limitation.

2.3.2 Bayesian Estimation with MCM

The use of Bayesian statistics has been controversial. Two most common criticisms against Bayesian approach are: 1) Imposing a probability distribution over a parameter is unreasonable because parameters are fixed. 2) Including a prior into the model may introduce subjective judgement and thus bias the results (Gill, 2014; Lynch, 2007). Since the mid-1990s, there has been an explosion in advances in Bayesian statistics due to pragmatic reasons. Many research questions in social and behavioral science readily lend themselves to a Bayesian approach. More importantly, the availability of Bayesian computational packages increases the ease of using sampling methods to estimate model parameters.

Within the frequentist framework, a parameter of interest is assumed to be fixed, meaning there is only one true population parameter. The Bayesian paradigm offers a very different perspective of hypothesis testing in that all unknown parameters are uncertain and viewed as random variables. Therefore, Bayesian statistics typically involves the use of probability distributions rather than point probabilities (Lynch, 2007; Schoot, Kaplan, Denissen, Asendorpf, Neyer, & Aken, 2014). Three ingredients of Bayesian statistics include prior distribution for the parameter, information provided by the data, and posterior inference. The relationship of these three components are represented through the Bayes' theorem:

$$f(\theta|data) = \frac{f(data|\theta)f(\theta)}{f(data)}, \tag{13}$$

where $f(data|\theta)$ is the sampling distribution and is proportional to the likelihood function, $f(\theta)$ is the prior distribution for the parameter, and $f(data)$ is a scalar. As shown in equation (13), the posterior distribution for the parameter θ depends on the prior probability distribution for θ weighted by the probability of the data given different values of θ.

The estimation of parameters using Bayesian approach is challenging because it derives a posterior distribution rather than a point estimate for the parameter (Fox, 2010; Lynch, 2007). Unlike the classical likelihood-based approach, the Bayesian method uses sampling methods to generate samples from the posterior distribution and then uses these samples to approximate the integrals of interest to help summarize the posterior distribution.

In general, the development of MCMC sampling methods and the growth in computing capacities have made Bayesian statistics more feasible and increased the popularity of Bayesian estimation (Brooks, 1998; Fox, 2010; Lynch, 2007). TheMCMC methods were popularized by a paper by Gelfand & Smith (1990), which discussed a class of algorithms for sampling a probability distribution based on constructing a Markov chain. Each step in the chain constructs an empirical distribution, and the chain converges through a certain number of steps to an equilibrium distribution which is the posterior distribution. Specifically, "Markov chain" refers to the process of sampling a new value for the parameter θ given its immediate predecessor θ^{-1}, and "Monte Carlo" refers to the random simulation process. Monte Carlo integration is used to approximate an expectation by using the Markov chain samples:

$$\int_S \ g(\theta)p(\theta)d\theta \cong \frac{1}{n}\sum_{t=1}^{n} g(\theta'). \tag{14}$$

The features of MCMC samplings include: 1) It facilitates sampling from complex distributions and handles multivariate densities; 2) It moves throughout the entire space of a posterior distribution. In particular, the Bayesian MCMC approach is favored by its flexibility. Unlike frequentist procedures relying on normality assumptions and asymptotic arguments, MCMC techniques can handle complex data, such as data with multilevel correlation structures or data on different measurement scales for different test items, by fitting a broader variety of models. In addition, Bayesian inference via MCMC is unbiased which has no requirement for the minimum sample size. The implementation of Bayesian methods was shown to be sensitive to some conditions. For example, prior distributions are required to be specified for unknown parameters, and thus the choice of prior often affects the final inference (Lambert, Sutton, Burton, Abrams, & Jones, 2005; Turner, Omar, & Thompson, 2001). Therefore, it was suggested to compare the marginal prior with the posterior distributions or compare the posterior results over a small number of prior variations (Müller, 2012). The other critical issue in the use of MCMC methods is to address the convergence problem by applying appropriate diagnostic tools. It was suggested to combine a number of strategies to reach reliable diagnosis of convergence, such as applying diagnostic procedures to a small number of parallel chains, monitoring autocorrelations and crosscorrelations, as well as modifying parameterizations or sampling algorithms appropriately (Brooks & Gelman, 1998; Cowles & Carlin, 1996). In addition, both the within-chains movement (i.e., check the trace plots or time series summaries) and the between-chains movement (i.e., check the impact of starting points on different chains) should be studied to monitor convergence (Gelman & Shirley, 2011). Diagnostic criteria to evaluate chain convergence include time-series plots, autocorrelation plots, density plots, and Gelman-Rubin statistic R (Brooks & Roberts, 1998; Cowles & Carlin, 1996). MCMC sampling can be conducted using different algorithms, and the ones most broadly used in simulation studies are the Metropolis-Hastings algorithm and Gibbs sampling. In addition, the Gibbs sampler can be used with certain component conditional distributions sampled through the Metropolis-Hastings algorithm, known as the Metropolis-Hastings within Gibbs sampling.

2.3.2.1 Metropolis-Hastings (M-H) Algorithm

The M-H algorithm is the most popular algorithm for MCMC sampling (Chib & Greenberg, 1995; Lynch, 2007; Patz & Junker, 1999b; Roberts & Smith, 1994). The iterative process is summarized as the following steps where i indexes the iteration count:

1) Choose a starting point $\theta^{i=0} = S$ for which $f(\theta|y) > 0$);
2) Draw a candidate parameter θ^* from a proposal/jumping distribution $\alpha(.)$, which can be symmetric or asymmetric, $\theta^* \sim \alpha(\theta^*|\theta^i)$;

3) Compute the acceptance ratio $r(\theta^{*}, \theta^{i-1}) = min\left\{\frac{f(\theta^{*}|y)\alpha(\theta^{i-1}|\theta^{*})}{f(\theta^{i-1}|y)\alpha(\theta^{*}|\theta^{i-1})}, 1\right\}$;

4) Sample *u* from a uniform distribution *U* (0, 1);

5) If $r > u$, then $\theta^{i} = \theta^{*}$, otherwise, $\theta^{i} = \theta^{i-1}$;

6) Set $i = i + 1$ and return to step 2 until enough draws are obtained. Otherwise, stop.

The M-H algorithm is simple. It uses a proposal distribution to sample a candidate value of the parameter given its current value, therefore, it requires a careful design of the proposal distribution. A poorly chosen proposal distribution would lead to low acceptance rate or slowly moving Markov chain and low efficiency of the Monte Carlo sampling.

2.3.2.2 Gibbs Sampler

Gibbs sampler is a special case of the M-H MCMC sampler using the ordered sub-updates (Alber, 1992; Albert & Chib, 1993; Fox, 2010; Lynch, 2007; Patz & Junker, 1999b). The proposal distributions match the posterior conditional distributions, and thus all the proposals are accepted, meaning that the acceptance ratio always equals to one. The iterative process is summarized as the following steps, where *i* indexes the iteration count and the parameter θ^{i} may be multidimensional or univariate partitioned into *p* subvector components:

1) Assign starting values to the parameter vector, $\theta^{i=0} = S$;

2) Sample $(\theta_2^{i+1}|\theta_1^{i+1},\ \theta_3^{i} \ldots\ \theta_p^{i}) \ldots (\theta_p^{i+1}|\theta_1^{i+1},\ \theta_2^{i+1} \ldots\ \theta_{p-1}^{i+1})$ and use the joint distribution of θ as a simulated value or an updated value from the posterior of θ;

3) Return to step 2 until obtaining a simulated sample;

4) Converge to the joint posterior as the equilibrium distribution.

The major feature of the Gibbs sampler is to reduce a multidimensional parameter into blocks and sample each block given the most recent values of other blocks. This allows a complex high-dimensional problem to be simplified and solved in low-dimension. However, the Gibbs sampling has several limitations. For example, the conditional distribution for each random variable in the model may not be derived from the posterior joint density function. It could also be that the conditional distribution has an unknown form, making it impossible to draw samples from it. In certain cases it takes a long time for the Gibbs sampler to move through all the regions of the density. As a result, Gibbs sampling could be inefficient and display slow "mixing".

2.3.2.3 Metropolis-Hastings within Gibbs (MHwG) Sampling

The M-H algorithm can be used within the Gibbs sampler, and the idea is to retain sequential sampling while sampling the conditional distribution via M-H steps (Merkle, 2005; Patz & Junker, 1999a). The iterative process is outlined as follows:

1) Choose starting values, $\theta^{i=0} = S$;
2) Draw a candidate parameter θ_1^* from a proposal distribution $\alpha(\theta_1|\theta_1^{i-1})$;
3) Accept θ_1^* with the acceptance probability $min\left\{\frac{f(\theta_1^*|\theta_2^{i-1},\theta_3^{i-1}\ldots\theta_p^{i-1})\alpha(\theta_1^{i-1}|\theta_1^*)}{f(\theta_1^{i-1}|\theta_2^{i-1},\theta_3^{i-1}\ldots\theta_p^{i-1})\alpha(\theta_1^*|\theta_1^{i-1})},1\right\}$;
4) Sample a value $\theta_1^i = \theta_1^*$. If θ_1^* is not accepted, set $\theta_1^i = \theta_1^{i-1}$;
5) Repeat steps 2-4 for the rest of the parameters θ_2, θ_3, …, θ_p.

As a hybrid algorithm, the MHwG sampling lessens the difficulty of specifying high-dimensional candidate distributions in the M-H algorithm as well as the difficulty of obtaining the posterior distribution for each parameter in the Gibbs algorithm.

2.3.2.4 Use of MCMC Estimation in Mixture IRT Models

The application of MCMC estimation in psychometric models was popularized by the work of Patz and Junker (1999a, 1999b), where the use of MCMC based on M-H sampling was demonstrated in 2PL IRT models. A review of the literature found that the MCMC estimation has been used in IRT-based models such as polytomous-ordered data (Patz & Junker, 1999a), rater effects and missingness (Patz & Junker, 1999a), nominal data (Wollack, Bolt, Cohen, & Lee, 2002), testlets (Bradlow, Wainer, & Wang, 1999), multilevel IRT models (Fox & Glas, 2001), and mastery classification (Janssen, Tuerlinckx, Meulders, & De Boeck, 2000).

Like IRT models, LCA assumes both the latent and the observed variables to be discrete. The use of MCMC in LCA is advantageous for estimating all the parameters simultaneously while accounting for the uncertainty. The examples of MCMC estimation in mixture modeling include model comparison and diagnosis (Garrett & Zeger, 2000), model selection (Carlin & Chib, 1995), jump diffusion sampling (Phillips & Smith, 1996), Gibbs sampler (Escobar & West, 1995), inequality and equality constrainted LC models (Hoijtink, 1998), as well as multilevel LC models (Vermunt, 2008). Moreover, MCMC has been shown to be useful in estimating models with covariates (Chung, Flaherty, & Schafer, 2006).

The MCMC methods have been used to deal with empirical data using mixture Rasch models, mixture Rasch models with a covariate, and mixture 3PL models as a way of estimating latent class membership, item and ability parameters, as well as mixing proportions. These studies addressed issues related to DIF item detection, guessing behavior, and test speediness.

The general idea of using MCMC for estimating mixture IRT models is to sample the class membership parameter for each examinee *i* along his/her continuous latent ability θ_i at each stage of the Markov chain, and then sample the parameters that characterize each distribution in the mixture within each latent class based on the sampled class parameters (Bolt, Cohen, & Wollack, 2001, 2002). The procedure of MCMC estimation for a mixture Rasch model is summarized below, where the subscripts *h, g, i* denotes examinee, latent class, and item respectively:

Step 1: Sample the class membership g (g = 1, 2, …, G) for each examinee h;

Step 2: Sample the latent ability θ_{gh} for each examinee h within class g;

Step 3: Sample the Rasch difficulty parameter b_{ig} of item i for class g;

Step 4: Sample the mixing proportions $\pi_g = (\pi_1, \pi_2, \dots, \pi_G)$ such that $\sum_{g=1}^{G} \pi_g = 1$;

Step 5: Sample the ability means μ_g and precisions σ_g for each class g.

Prior distributions need to be specified to estimate the posterior distribution of each parameter. Consider the two-class solution as an example (g = 2), below are the commonly used prior distributions in mixture Rasch modeling, while σ_g is recommended to be fixed at 1 for both groups or assigned a Gamma prior:

$g_h \sim Bernoulli(\pi_1, \pi_2), h = 1,2, \dots, N;$

$\theta_{gh} \sim Normal(\mu_g, 1), h = 1,2, \dots, N, g = 1,2;$

$b_{ig} \sim Normal(0,1), i = 1,2, \dots, I, g = 1,2;$

$\pi_g = (\pi_1, \pi_2) \sim Dirichlet(0.5, 0.5), g = 1,2;$

$\mu_g \sim Normal(0,1), g = 1,2;$

$\sigma_g \sim Gamma(1,1), g = 1,2;$

The early work focused on examining the effect of test speededness on item parameter estimates and scale stability by using mixture Rasch models with MCMC estimation (Bolt, Cohen, & Wollack, 2002; Cohen, Wollack, Bolt, & Mroch, 2002; Wollack, Cohen, & Wells, 2003). In these studies, MCMC was used to handle ordinal constraints on the model parameter b_{ig} in the mixture Rasch model such that the item difficulty estimates were constrained to be higher on one class than the other class, $b_{i1} > b_{i2}$. To ensure adequate convergence, prior distributions for different model parameters were specified. Given the computational speed of the MCMC estimation, a two-stage approach was used to expedite the MCMC algorithm through fixing item parameter values at their estimates from the previous run and then the MCMC algorithm on the new data was re-run.

Along this line, in a study about test-taking behavior, the mixture Rasch IRT model with MCMC estimation was used to fit test data by incorporating item response time parameter (Meyer, 2010). In order to estimate this additional parameter, the means of item response time, vague conjugate priors were used and order constraints were imposed such that the solution behavior class had larger item response time means than the rapid-guessing class. The item response time and model parameters were estimated using MLE and JMLE respectively to justify the efficacy of MCMC. The correlation matrix of item difficulty values for each class and estimation method was also examined.

In Cohen and Bolt’s study on DIF (2005), MCMC with Gibbs sampling was used to estimate the class membership of examinees using a mixture 3PL model while fixing the item parameter estimates obtained from a multigroup MULTILOG procedure. Three parameters were estimated including θ_{gh}, π_g, and μ_g. This study demonstrated that the two latent groups did not match the gender makeup of the sample. In other words, the latent classification did not agree with the manifest variable classification, and the mean difference between the two latent classes was larger. The findings led to later research investigating the cause of DIF based on mixture IRT models.

Rather than classifying persons into different latent groups, Frederickx, Tuerlinckx, De Boeck, and Magis (2010) employed a normal mixture distribution to model item random effect with two components representing DIF item class versus non-DIF item class. Based on theoretical considerations and practical constraints, the MCMC algorithm was used with both vague prior (i.e., uniform, inverse gamma, and truncated normal for the standard deviation parameters) and informative priors (i.e., $N(0,1)$ for the mean parameters). In addition, a convergence measure $\hat{R}$ was calculated to represent the convergence quality of the Markov chains, which is approximately the square root of the ratio of the between-chain variance to the within-chain variance.

Dai (2013) and Li, Jiao, & Macready (2015) investigated the performance of the mixture Rasch model with covariates under MCMC estimation Gibbs sampler. Considering the stability of MCMC estimation, Dai (2013) recommended to monitor the critical steps of the MCMC algorithm to identify non-converged MCMC chains, even though this did not guarantee good estimation of the parameters due to potential label switch issues. Two chains with different start values were considered to be merged properly only if they merged into the same stable region and provided reasonable values for the parameter estimates. This was checked by looking at the history graph. Li, Cohen, Kim, and Cho (2009) compared model selection methods in dichotomous mixture IRT models. The proposed indices were specifically for a Bayesian solution, including the pseudo-Bayes factor (PsBF), the deviance information criterion (DIC), the posterior predictive model checks (PPMC), and the MCMC estimation version of AIC and BIC. Convergence diagnostic was also performed to determine how many iterations were burn-in or could be used to estimate the posterior distributions. The convergence issue was also addressed by examining the covariate effects on mixture Rasch models within the Bayesian framework (Li, Jiao, & Macready, 2015). It was stated that the within-chain label switching coupled with the non-systematic fluctuations in different chains and the complexity of model may cause poor mixing or non-convergence.

2.3.3 Estimation Efficacy of Mixture IRT Models

Maximum likelihood and Bayesian analysis with MCMC sampling are the most frequently used estimation methods in mixture IRT modeling. There were studies supporting the appropriateness of the MCMC algorithm in generating reasonable class classification accuracy and item parameter recovery. However, a review of mixture IRT models used in practice suggested that relatively little research was conducted to directly compare between the estimation procedures within the mixture IRT framework (Cho, Suh, & Lee, 2015). The evidence for the estimation efficacy of mixture IRT models mainly comes form the simulation studies on the performance of the MCMC estimation and its efficiency compared to MLE.

In a preliminary study, Bolt, Cohen, and Wollack (2001) investigated individual differences in the selection of response categories for multiple-choice questions by fitting a mixture nominal response model via MCMC estimation. Datasets were generated for 12 five-category items under the two-class condition. The accuracy of parameter estimates was measured by the Root Mean Square Error (RMSE), and the classification accuracy was measured by the proportion of examinees correctly classified into the class where they were simulated, termed as hit rate or percentage of correct identification. By using the MCMC adaptive rejection sampling (ARS) that applies to log-concave conditional distributions, two item category parameters were well recovered across all the simulation conditions, whereas other model parameters were affected by the manipulated factors as expected. Specifically, the smaller between-class difference led to poorer classification accuracy and poorer parameter recovery within each class. The mixing proportions of the classes also had an effect on classification accuracy. This study provided initial support for the use of MCMC estimation in mixture IRT models.

In a simulation study identifying the optimal model selection indices using the Gibbs sampler (Li et al., 2009), the recovery analysis used the same accuracy measures of classification and parameter estimates as above. The latent group membership was well recovered (i.e., all greater than 80%), while the percentage of correct classification was reduced as the model became more complex (i.e., lowest percentage for the 3PL model). However, the recovery of group membership improved as the number of items increased regardless of model complexity. The recovery of the item parameters was reasonable as well. It was found that the recovery of the item discrimination and difficulty parameters was especially worse as the number of latent groups increased, whereas the recovery of the guessing parameter did not depend on the number of simulated latent groups, test length, or sample size.

The use of the MCMC algorithm was expanded to the mixture multilevel IRT models to explain the DIF by detecting and comparing the characteristics of the latent groups (Cho & Cohen, 2010). Two latent classes were generated at both the student and the school level, and data was fitted using the 1PL model. The study showed that the recovery of the group membership was good at both the student and school level. The item difficulty parameter was also well recovered, which was indicated by RMSE and bias. This study added evidence to the efficiency of MCMC estimation in modeling mixture multilevel item response data.

More recently, several studies that compared the MCMC approach with MLE further claimed the superiority of MCMC technique in certain conditions (Finch & French, 2012, Cho, Cohen, & Kim, 2013). MCMC has been proved to be useful for complex mixture IRT models without a requirement for the integration of the likelihood function, which could be difficult to be achieved in MLE when many parameters were involved (Junker, 1999).

Finch and French (2012) compared the performance of the marginal MLE and MCMC estimation in classification accuracy and parameter estimation bias based on mixture 1PL and 2PL IRT models for dichotomous item response data. Several conditions were manipulated including the number of latent classes (2, 3, 4), the number of items, total sample size, and group size/ratio. Overall, the MCMC method led to a higher classification accuracy rate (uniformly greater than

90%) across all the levels of the manipulated factors, and the method of estimation interacted significantly with each of these factors. For example, the gap in classification accuracy between the two methods narrowed as more items were included, and the performance of the MCMC estimation deteriorated as the group size became unequal. In addition, the MCMC method produced smaller bias in the estimation of the discrimination parameter regardless of the number of items and was not affected by group size. Moreover, the coverage rates for MCMC estimates were much higher (near 1.0) than those for the MLE estimates. It was concluded that the MCMC estimation provided better recovery of the group membership across conditions and more accurate parameter estimates when the sample size and the number of items were small. Instead, when more items were included, the MLE method produced smaller confidence intervals and thus more accurate parameter estimates.

Cho et al (2013) examined the impact of priors on the probabilities of mixtures, label switching, model selection, and metric anchoring by using the mixture Rasch model. The simulation analysis revealed good recovery of class membership under two-class conditions regardless of priors, whereas the recovery of item parameters depended on such factors as the number of latent classes simulated, test length, and sample size. The focus of this study was not on comparing estimation algorithms, however, as a part of the study the MCMC algorithm, implemented in WinBUGS, was compared with the CMLE based on three different computer programs (WINMIRA, LatentGOLD, and Mplus) using an empirical dataset from an 18-item math test. To measure the agreement between the MCMC and MLE item difficulty estimates, correlations for all pairs of estimates obtained from all four computer packages were computed, which were higher than 0.99. Likewise, the Kappa coefficients were computed as a way of describing the agreement in group membership identification, which also suggested good agreement except for a few pairs. Finally, the covariate effects, measured by the slope coefficient estimates, were found similar across the computer packages.

2.4 ISSUES IN ESTIMATION FOR MIXTURE IRT MODELS

As a combination of the finite mixture model and the IRT model, mixture IRT models are subject to some common model identification issues, and the major two were referred to as the scale indeterminacy or metric identification and label switching (Baker & Kim, 2004; Jasra, Holmes, & Stephens, 2005). Previous studies addressed the solutions for each issue within the mixture IRT framework (i.e., Choi, 2014; Dai, 2013; Paek & Cho, 2015), which are discussed below.

2.4.1 Metric Identification

Since mixture IRT models categorize examinees into different latent classes and allow ability and item parameters to be different, a main goal of using mixture IRT models is to compare item profiles across latent groups and characterize examinees from different latent groups. To ensure comparability, a common scale/metric needs to be established across latent groups before making any comparison.

This metric identification issue derives from the IRT portion of the mixture IRT models. Scale indeterminacy, also known as metric indeterminacy or metric identification issue has been known as a property of IRT, which refers to the arbitrariness in the choice of the origin and scale of the metric. The same distance between the person and item locations (i.e., $\Delta = 1$) leads to the same probability of correct response. As a result, multiple ability and item parameters lead to the same response probability and the metric is not absolute or unique (De Ayala, 2013). To estimate person and item parameters the metric needs to be fixed or anchored to a certain origin.

Similar to the strategies used with IRT models, researchers recommended three methods to place person and item parameters on a common scale across different latent groups. The first method is known as concurrent calibration, where a set of anchor/class invariant items are used to anchor the metric across latent classes (Von Davier & Yamamoto, 2004). The item parameters of anchor items are fixed to the same values in order to establish identifiability and comparison across all latent groups. For example, in an empirical study on DIF analysis based on a mixture 3PL model with a covariate, researchers identified the items that functioned the same across latent groups using the likelihood ratio test, and then constrained their discrimination, difficulty, and guessing parameters to be equal to anchor the metrics of latent groups such that the item parameter estimates were comparable across groups (Choi, Alexeev, & Cohen, 2004). The challenge of this method is to identify appropriate anchor items.

The second method, known as the equality constraint, is to arbitrarily select a latent class as the reference group and fix its ability distribution as N(0, 1), and then the estimation of the model parameters for other groups can be located relative to the scale of the reference group (Cho, Cohen, & Kim, 2014). This person centering method is straightforward, however, some researchers argued that imposing constraint on ability distribution alone does not guarantee a common scale between latent groups when the ability distributions of the latent groups are not identical or item profiles are different (Paek & Cho, 2015).

The third method, known as the item centering method, was proposed by Rost (1990) based on the mixture Rasch model, which is to constrain the sum of the item difficulty $\sum b_{jg}$ to be zero for each latent group. For example, in a study evaluating the performance of the mixture Rasch model under the MCMC estimation, Dai (2009) assigned the items with large DIF effect sizes across the latent groups, and then gave opposite signs to the items with relatively small DIF effect sizes. By this way the sum of item difficulties $\sum b_{jg}$ was set to zero for each latent group.

Choi (2014) compared these three methods (item anchoring, person centering, and item centering) in establishing a common metric between latent groups by using three different mixture IRT models (mixture Rasch, mixture 2PL, and mixture 3PL). Factors that were manipulated included sample size (600 v.s. 2,400), test length (20 v.s. 40), and the number of latent groups (1-, 2-, and 3-group). The recovery results showed that the constraint type had no significant effect on the recovery of item difficulties as was indicated by RMSE. It did not affect the identification of the latent group membership either. In addition, the mixture 2PL model had the best recovery of the latent group membership compared to the mixture Rasch or the mixture 3PL model, which also had relatively small bias values for the item discrimination estimates. Lastly, the correlations between item difficulty and item discrimination were moderately high to high for all three constraints. It was suggested by the author that any of these constraints would be useful for the estimation of the mixture Rasch or mixture 2PL model, whereas for the mixture 3PL model the item anchoring constraint would be recommended.

2.4.2 Label Switching

Label switching is a well-known estimation issue in mixture modeling, which is typically unavoidable and is usually a concern in the Bayesian estimation of mixture IRT models. It describes the invariance of the likelihood under relabeling of the mixture components (Redner, 1984; Stephens, 2000), that is, the likelihood is the same for all the permutations of the component-specific parameters.

Two types of label switching is identified in the literature: the within-chain label switching and the between-chain label switching (Cho, Cohen, & Kim, 2006, 2010, 2013; Cho, Suh, & Lee, 2015; Dai, 2013; Finch & French, 2012; Li et al., 2008). The within-chain label switching occurs across iterations or repeated sampling from the posterior distribution within a single MCMC chain. When it occurs, the labels of the latent classes switch at each iteration within a single analysis and thus the interpretation of the meanings of the latent classes would be distorted (i.e., non-unique labeling of a latent class). This type of label switching typically leads to $k!$ symmetric modes of the parameter posterior distribution for a mixture model with k components/subspaces. In this case, the MCMC sampler would visit one of the $k!$ modes only and fail to thoroughly travel through the distribution surface, resulting in poor or unreliable parameter estimation (Frühwirth-Schnatter, 2006; Jasra, Holmes, & Stephens, 2005; Marin, Mengersen, & Robert, 2005). The within-chain label switching can be detected by observing if multiple modes are present in the posterior distribution of the parameter.

The between-chain label switching occurs across different chains or the replications of a simulation process. When this occurs, the labels of the latent classes change from one replication to another (i.e., the high ability group in on run becomes a low ability group in the next run). This type of label switching is found with both the Bayesian estimation and MLE. However, it has not received as much attention for frequentist mixture models as for Bayesian mixture models. The between-chain label switching may cause confusion when using empirical dataset, but can be easily detected and relabeled by comparing the parameter estimates of each replication with the generating parameters when using simulated data. While for empirical data, the between-chain label switching can be detected by comparing the estimates between two runs and between the latent groups. One of the runs can be set as a reference, and then other runs can be compared with the reference run to see if the parameter estimates are consistent or if relabeling is needed (Cho et al., 2015).

Several solutions were proposed to solve the label switching issue. The first solution is straightforward and was recommended by Dai (2013) for mixture Rasch models with covariates. The history plots of mixing proportions can be examined to determine what type of label switching occurs. The between-chain label switching can be simply relabeled, while the within-chain label switching can be excluded from the analysis.

The second solution is to impose parameter constraint such as $\pi_1 < \pi_2 < \cdots < \pi_k$ and $\mu_1 < \mu_2 < \cdots < \mu_k$. However, this strategy was considered inefficient in that one can find permutations $\rho_1 < \rho_2 < \cdots < \rho_N$ such that the parameter constraint is satisfied by the permuted sample $\rho_1\left(\theta^{(1)}\right) \ldots \rho_N\left(\theta^{(N)}\right)$ (Dellaportas & Stephens, 1996; Diebolt & Robert, 1994; Richardson & Green, 1997; Stephens, 1997).

The third solution is known as the relabeling algorithm, which basically is the k-means type clustering of MCMC samples (Celeux, 1998; Celeux, Hurn, & Robert, 2000; Stephens, 1997, 2000). The basic idea is to minimize the posterior expectation of some loss function, and the points closest to the current cluster means at each iteration of the $k!$ permutations are selected. According to Stephens (2000), one way of measuring loss is the Kullback-Leibler divergence of the classification probabilities $p(\theta)$ from the true distribution on clustering.

The fourth solution is to use random permutation samplers (RPS), which was proposed by Frühwirth-Schnatter (2001). One of the $k!$ permutation label order $\rho(1) \ldots \rho(k)$ is drawn to substitute the current MCMC sample's component order $\theta(1) \ldots \theta(k)$. The substituted component order would be $\theta_{\rho(1)} \ldots \theta_{\rho(k)}$. This ensures the sampler visits all $k!$ symmetric modes.

The fifth solution is to apply the ascending algorithm (ALG) based on the posterior model labeling, which was proposed by Yao and Lindsay (2009). Each MCMC sample is used as the starting point for an ascending algorithm, and the label is assigned based on the mode where the algorithm converges. All other permuted maximal modes would have clear labels referring to the maximal modes. This algorithm has the computational advantage that it does not require the comparisons to $k!$ permutations when assigning a label except for minor labels.

All the above solutions are targeting at the labeling switching issue in Bayesian mixture models. The last solution, known as the complete likelihood based labeling (COMPLH), was proposed by Yao (2015) to deal with the label switching problem for frequentist mixture models. The label is found through maximizing the complete likelihood of the observed and latent variables with respect to the permutation $L(\hat{\theta}^{\rho}|x, z)$. This complete likelihood $L(\hat{\theta}^{\rho}|x, z)$ is not invariant to the permutation of component labels because the latent variable brings in information for labeling and helps break the permutation symmetry of the mixture likelihood.

2.5 LATENT DIF ANALYSIS USING MIXTURE IRT MODELS

The concept of measurement invariance (MI) was introduced by Mellenburgh (1989), and was defined as the parameters of a model independent of group membership (Meredith, 1993). A fundamental principle of measurement is that a scale is measuring the same trait across two or more subpopulations of a sample. The violation of this property implies that individuals with identical latent traits but from different groups score differently on a test item. When measurement non-invariance occurs, it is difficult to tell whether the test score differences can be attributed to the construct being measured or the differential functioning of the test items across groups. MI can be tested by examining the equity of item response functions based on IRT models (Thissen, Steinberg, & Gerrard, 1986), that is, to determine if the conditional probability of observing a response pattern is invariant across groups given the ability parameter θ. The violation of MI is commonly referred to as DIF in the IRT literature (Drasgow, 1984; Maurer, Raju, & Collins, 1998; Meade & Lautenschlager, 2004; Raju et al, 2002; Reise et al, 1993).

Traditionally, DIF is defined in terms of manifest variables such as gender, culture, ethnicity, age, etc. In recent years researches started investigating DIF across latent groups, which was argued to be a better way of discovering the true sources of the group differences. This section focuses on the description of latent DIF and how it is identified within the mixture IRT framework.

2.5.1 DIF and Latent DIF

DIF is broadly defined as the psychometric difference between the item functions of two examinee groups (Dorans & Holland, 1993). Items are considered to provide equivalent measurement if the item parameters remain invariant across two populations (Raju et al, 2002; Reise et al, 1993):

$$a_i = a_i', \tag{15}$$

$$b_i = b_i', \tag{16}$$

where the prime denotes the second population. Two types of DIF defined in the literature are uniform and nonuniform DIF. Uniform DIF refers to no interaction between the ability level and the group membership. In this case the item response functions (IRFs) are parallel with one group consistently favored or disadvantaged to the other group. The non-uniform DIF occurs where there is an interaction between the ability level and the group membership. In this case the IRFs cross at a certain ability θ point, suggesting that the differences between two groups differ in magnitude and direction across the ability scale (Mellenbergh, 1989; Swaminathan & Rogers, 1990). Each type of DIF can be represented graphically as follows:

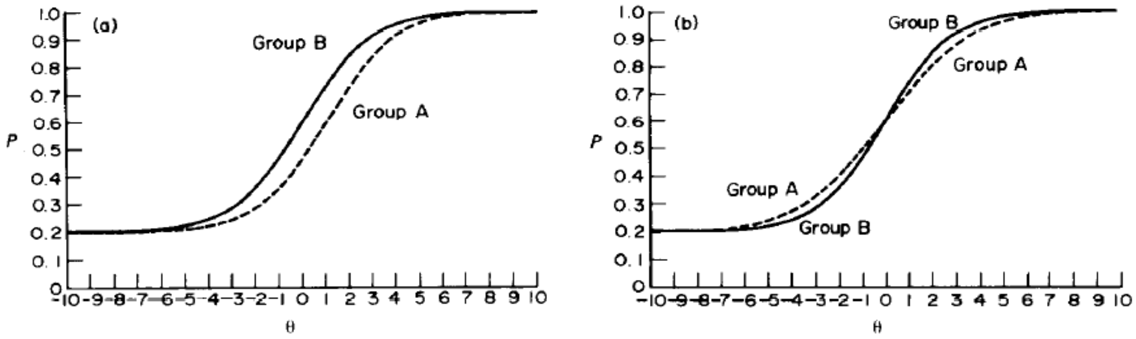


Figure 2.2. Examples of Uniform and Nonuniform DIF (Mellenbergh, 1989)

The cause of uniform DIF is the shift of the *b* parameter, while the cause of nonuniform DIF is the shift of the *a* parameter and possibly the *b* parameter. The two types of nonuniform DIF are further distinguished. The one with IRFs crossing within the range of ability level (typically from -3 to +3) is analogous to a disordinal interaction in ANOVA, and is termed nondirectional DIF. In contrast, the one with nonparallel IRFs but crossing outside the range of ability level is analogous to an ordinal interaction, and is termed unidirectional DIF (Li & Stout, 1996; Narayanon, & Swaminathan, 1996; Swaminathan & Rogers, 1990).

Classifying examinees into groups (i.e., male group v.s. female group) is based on manifest grouping variable (i.e., gender) and is determined prior to the DIF analysis. However, in the situation where items function the same way for an unknown homogenous group, DIF may still exist due to certain unobserved features of examinees such as unmeasured educational background, personality traits, attitudes, etc. As opposed to manifest DIF, this type of DIF is defined across unknown groups. Several early studies showed that the latent class membership did not necessarily overlap with the manifest class membership, and items functioning differentially against one manifest group (i.e., female group) may not against all the members in that group (i.e., all the females). Furthermore, examinees' ability differences based on manifest groups were not consistent with the ones based on latent groups (Cohen & Bolt, 2005; De Ayala et al., 2002). Researchers argued that understanding latent DIF contributes better to the explaining of the cause of examinees' response differences on DIF items because it taps the nuisance dimensions that account for the mechanism that gives rise to DIF (Cohen & Bolt, 2005; De Ayala et al., 2002).

2.5.2 Analysis of Latent DIF

Benefitting from the capability of its mixture modeling portion in uncovering latent homogeneous groups, the detection of DIF using mixture IRT models has received increased attention. Based on an empirical dataset, Cho et al (2015) identified the common procedure of detecting DIF with mixture IRT models: 1) Select the best fitted model referring to model fit indices such as Bayesian information criterion (BIC) and determine the number of latent classes; 2) Identify the latent group membership for each person in terms of the selected measurement model; 3) Conduct DIF analysis to identify items that characterize examinees differentially across the latent groups (Cho et al., 2015). It is noted that latent DIF can be detected using the same methods developed for manifest DIF including the non-parametric and parametric procedures.

When covariates are incorporated into the models, extra steps may be added to the above procedure depending on the way covariates are associated with the latent class membership. Three approaches that describe the way covariates are incorporated are known as the one-step, two-step, and three-step approach. The one-step approach establishes the relationship between the covariates and the latent class membership through a one-step process, that is, to integrate the covariates into mixture IRT models directly (Dai, 2009; Li et al., 2015). By using the one-step approach, the adding or removing of a covariate would result in both the prediction and measurement models to be re-estimated and the latent classes to be re-defined (Vermunt, 2010). The two-step approach estimates item parameters and examinees' latent class membership without covariates in the first step, and then establishes the connection between the latent class membership and the evident variables of interest in the second step to explain the sources of latent DIF (Cohen & Bolt, 2005; De Ayala et al., 2002; Samuelsen, 2005; Van Nijlen & Janssen, 2008). The three-step approach was proposed by Vermunt (2010), which corrects for the uncertainty in the identification of latent class membership. The latent class model is estimated in the first step using the entire sample data. In the second step, the most likely class variable is created based on the latent class posterior distribution obtained from the first step. In the third step, this most likely class variable is used to estimate the model. The present study adopted the one-step approach because the goal is to examine the covariate effect on the classification of latent groups and the estimation of model parameters simultaneously.

A review of the literature found that several simulations studies on latent DIF provided evidence for the effectiveness of mixture IRT models (primarily the mixture Rasch model) and discovered the factors/conditions that may influence the detection of latent DIF. Samuelsen's work (2005) was one of the early studies that investigated the application of mixture Rasch models on DIF detection. In this study, data was simulated on a 20-item test. The manipulated factors were sample size (500 v.s. 2000), manifest proportions (50/50, 80/20), the overlap between the manifest groups and the latent classes (100%, 90%, 80%, 70%, 60%), number of DIF items (2, 6, 10), effect size of DIF (0.4, 0.8, 1.2), and the ability distributions within the latent classes (N(0, 1), N(-1, 1)). MCMC estimation was used and the constraint of $\sum b_{jg} = 0$ was applied to each latent class to solve the model identification issue. The background variables were selected as categorical and continuous covariates. It was found that increasing sample size increased the precision of DIF detection, especially when a categorical covariate was highly overlapped with the latent classes. Also, the high overlap reduced the standard errors of estimation, and the large sample size reduced the bias of the mean ability differences between the latent groups. Based on this work, Lu and Jiao (2009) found that the differences in mean ability between latent groups, large DIF, large number of DIF items, and equally split latent groups contributed to accurate identification of DIF items. In addition, the accuracy of class assignment was approximately 70% for the conditions where DIF items were most accurately identified.

Dai (2009) studied the role of covariates in mixture Rasch models by manipulating the manifest proportions (50/50, 30/70), latent proportions (15/85, 30/70, 50/50), number of items with large DIF effect sizes (6, 12), ability distributions within each latent class ($N(0, 1)$ v.s. $N(0, 1)$, $N(0, 1)$ v.s. $N(1, 1)$), and the association between the covariate and the latent classes indicated by odds ratio (1, 10, 50). Data were simulated on a 30-item test, and only one dichotomous covariate was used. The sample size was fixed to 1000. Similar to Samuelsen's study (2005), the number of latent groups was fixed to two. Bayesian MCMC estimation was used with 11 replications in each cell and the constraint of $\sum b_{jg} = 0$ was applied to each latent class. It was found that most the simulation cells could identify more than 50% of the DIF items with large effect sizes. The higher percentage of items with large DIF effect sizes led to better identification of the underlying DIF and better recovery of the latent structure. The focus of this study was on the role of the covariate. It was found that the latent class membership was better recovered when the manifest-latent relationship was moderate or strong.

Studies that explored the effectiveness of the mixture IRT models in revealing DIF between unobserved groups further highlighted the influence of sample size, correlation of manifest and latent variables, equality of the latent group size, number of DIF items, as well as group impact (DeMars & Lau, 2011; Maij-de Meij, Kelderman, & van der Flier, 2010).

3.0 METHODOLOGY

The main purpose of this study was to examine how well the mixture IRT model with covariate performed in identifying latent DIF. This chapter describes the study design, estimation methods, outcome measures, and solutions to the label switching problem. In addition, the simulation process and the validation of the simulation conditions are discussed to provide a protocol for future research.

3.1 DESIGN OF THE SIMULATION STUDY

This section describes the factors that were manipulated and kept constant in the current study. Given the complex nature of the models employed, several conditions were kept constant to make the simulation manageable. These constant factors are summarized in Table 3.1 below.

Table 3.1. Factors Kept Constant in Simulation

Constant Factors	**Values**
Test length	40
Number of latent classes	2
Sample size	2,000
Manifest proportion	50/50
Latent mixing proportion	70/30
Number of covariate	1

The number of test items was fixed to 40, which was more than the number of items used in previous studies (Dai, 2009; Samuelsen, 2005; Smit et al., 1999). This test length was selected to mimic the scenario of a typical large-scale state assessment. For example, a total of 43 multiple-choice questions were used in the 2016 Texas Assessment of Academic Readiness for Mathematics for students in 3rd grade. The 2016 New York State Common Core test for Mathematics included 44 multiple-choice questions for students in younger grades. Therefore, a test of 40 multiple-choice items reasonably resembles real practice in state assessment.

The number of latent classes was fixed to two, which was consistent with most of the previous studies. As was suggested by Li et al (2009), the recovery of item discrimination and difficulty parameters primarily depend on the number of latent groups simulated. When the number of latent groups increased, the recovery of these two parameters became worse regardless of the mixture IRT models fitted (1PL, 2PL, or 3PL). Therefore, a constraint of two latent groups ensures reasonable recovery of item parameters.

A total sample size for each replication was kept at 2,000. Some researchers used a relatively small sample size of 1,000 (Dai, 1999), some used a large sample size of 3,000 for each latent group which added up to 6,000 in total (DeMars, 2011), and some treated it as a manipulated factor taking the values of 1,000, 5,000, and 25,000 (Maij-de Meij et al., 2010). It was noted that a relatively small sample size was sufficient to ensure accurate estimation of model parameters when the mixture Rasch model was fitted. In a simulation study comparing the model selection indices based on three mixture dichotomous IRT models (1PL, 2PL, 3PL), sample size was manipulated to be 600 or 1,200 (Li et al., 2009). It was suggested that even for the 2PL or 3PL model that required the estimation for more parameters, the sample size of 1,200 still ensured satisfactory accuracy of item parameter estimates and recovery of the latent group membership (i.e., > 90% for the 2PL and 3PL models for a test of 30 items). In addition, the recovery of the latent group membership was not affected as the sample size increased. Furthermore, the focus of the current study is the role of a covariate in latent DIF detection and the impact of estimation methods on mixture IRT modeling. Therefore, keeping the sample size constant helps reduce the complexity of the study design.

The composition of the covariate was fixed to 50%: 50% to mimic the primary distribution of manifest groups in population such as gender and school location (i.e., urban v.s. non-urban). In contrast, the latent mix proportion was fixed to 70%: 30%, which was consistent with the findings from empirical studies. It was found that in testing practice examinees were rarely equally distributed among different latent groups. The 70/30 proportion was designed to represent the uneven distribution of the latent groups. Specifically, the 70% group represented the reference latent group, and the 30% group represented the focal latent group.

Lastly, one dichotomous covariate was included because the primary interest of this study is to examine the role of the joint relationship between the covariate and the latent group membership in the mixture IRT modeling of DIF. Therefore, more than one covariate was not considered.

Seven factors were manipulated including five between-replication factors and two within-replication factors. To better inform psychometric practice, the levels of each manipulated factor were chosen to resemble the real-world situation as much as possible, which is also consistent with the simulation conditions used in previous studies. These factors and their corresponding levels are summarized in Table 3.2 below.

Table 3.2. Factors Manipulated and Corresponding Levels in Simulation

Manipulated Factors	**Values**
Proportion of items with DIF	15%, 30%
DIF type	Uniform, Non-uniform
DIF size/magnitude (Δa or Δb)	0.5, 1
Group impact	N(0, 1), N(0, 1); N(1, 1), N(0, 1)
Strength of the relation between D_j and π_{jg} (OR)	2, 8
Analyzing model	No covariate, Covariate
Estimation method	MLE, Bayesian

In the real testing scenario, the number of DIF items was typically less than 30% (Güler & Penfield, 2009; Narayanan & Swaminathan, 1994; Puhan, Moses, Yu, & Dorans, 2009). Previous studies found that the number of items exhibiting DIF was between 15% to 30% (Hambleton, R. K., & Rogers, 1989; Raju, Bode, & Larsen, 1989). When a large number of DIF items existed, the test would probably measure two distinct constructs in the two groups (DeMars, 2011). Therefore, the use of 15% and 30% DIF items resembles previous DIF research. Two proportions of DIF items were considered in the present study, which corresponded to 6 and 12 items.

Two types of DIF were included in this study. The uniform condition meant that all the DIF items were uniform DIF items, which was consistent with previous simulation studies and the focus of DIF research. DIF research is related to fairness in testing. A focus on uniform DIF answers how test difficulty differentiates between groups and which group is unintendedly disadvantaged. Nonuniform DIF implies an interaction between the ability level and the group membership. Practically, uniform DIF occurs more often than non-uniform DIF in standardized tests (Narayanon & Swaminathan, 1996), but non-uniform DIF items are still identified (Hambleton & Rogers, 1989). Therefore, the other condition specified all the items to be non-uniform DIF items. When non-uniform DIF is present, item discrimination differs between the two latent groups, while item difficulties are constraint to be equal. The situation where all the DIF items are non-uniform was impractical, however, the effect of shifting item difficulties on DIF detection accuracy can be isolated and the capability of the mixture 2PL model in picking up non-uniform DIF items can be better assessed (Stark, Chernyshenko, & Drasgow, 2006).

The size of DIF indicates the magnitude of the differences in item difficulty or item discrimination between groups. Previous studies manipulated the size of uniform DIF, which was represented by the absolute value of the difference in difficulties between two latent groups, $\Delta b = |b_{j1} - b_{j2}|$. In present study, the difference between the reference latent group and the focal latent group was used as the magnitude of DIF. According to Zwick and Ercikan (1989), Δb less than 0.5 can be considered as negligible DIF, between 0.5 to 1 can be considered as moderate DIF, and greater than 1 can be considered as large DIF. The selected uniform sizes are consistent with previous studies (Dai, 2009; Penfield, 2001). Likewise, two levels of item discrimination difference were selected as a manipulation of the non-uniform DIF size. Previous study suggested Δa between 0.22 to 0.63 to be low non-uniform DIF and Δa between 0.78 to 1.54 to be high non-uniform DIF (Narayanon & Swaminathan, 1996). Therefore, the low and high magnitudes of non-uniform DIF also correspond to Δa = 0.5 and Δa = 1.

Consistent with previous studies (Dai, 2009; DeMars, 2011; Li et al., 2009; Li et al., 2015), two impact levels were selected by manipulating the differences in group abilities. For the no impact condition, both the latent groups have their abilities following the standard normal distribution N(0, 1). For the impact condition, the focal group had the ability following N(1, 1), while the reference group following N(0, 1).

Logit parameterization can be used to estimate the association between the observed variable and latent trait (Eid & Langeheine, 1999). Based on the logistic regression model, the odds ratio can be used to describe the relationship between a covariate and latent class membership. When the odds ratio equals one, it implies no effect of the covariate on latent membership. In the current study, low and high odds ratio values were specified to manipulate the relationship between the covariates and latent group membership. Referring to table 3.3, the odds ratio can be represented using the number of examinees in each cell. For example, the odds ratio of 2 is equivalent to 772 examinees classified into the reference latent group and the manifest group 1, 628 examinees into the reference latent group and the manifest group 2, 228 into the focal latent group and the manifest group 2, and 372 examinees into the focal latent group and the manifest group 2. These numbers of examinees in each cell are used in data generation to represent the relationship between the covariate and the latent variable.

Table 3.3. Representation of the Relationship between Covariate and Latent Groups

	Manifest Group					
	Group 1			**Group 2**		
Latent Group	OR = 2	OR = 8	Total (N)	OR = 2	OR = 8	Total (N)
Reference Group	772	628	1400	892	508	1400
Focal Group	228	372	600	108	492	600
Total (N)	1000	1000	2000	1000	1000	2000

The last two factors manipulated were the model and estimation method used. Models with and without a covariate were used to estimate each replication dataset, and the MLE and Bayesian estimation were also used to estimate each replication dataset. These two factors can be considered as the within-replication factors. Since the latent groups were generated by considering the relationship with the covariate, the mixture 2PL model with the covariate is considered as the correctly specified estimating model, and the model without the covariate is considered as the mis-specified estimating model.

In summary, the levels of each factor were selected to be consistent with previous studies and testing practice. Factors that were kept invariant included test length, number of latent groups, sample size, the distribution of the covariate, latent mixing proportion, and the number of covariates. Seven factors were varied including proportion of DIF items, DIF type, DIF size, group impact, and the relationship between the covariate and latent group membership, the data analyzing model, and the estimation method. Altogether, there were 2 × 2 × 2 × 2 × 2 × 2 × 2= 128 conditions. 50 replications were run for each of these conditions, and the total replications were 128 × 50 = 6,400.

3.2 PARAMETER ESTIMATION

Another main purpose of the current study is to investigate the effect of estimation method on the performance of mixture 2PL model on DIF detection. MLE and Bayesian estimation were used to estimate the models separately. Consistent with the literature (Finch & French, 2012), MML was selected and implemented in Mplus. Since MLE may mistakenly converge on the local maxima and lead to suboptimal parameter estimates, multiple random starting values were used by setting

“STARTS = 100 10” in Mplus (Cho et al., 2015). This meant that 100 random sets of starting values were used in the initial stage and 10 random values were used in the final stage of optimization. The use of more starting values improves the probability of obtaining optimal fit but increases the maximum number of iterations.

Bayesian estimates based on the mixture 2PL model were obtained using the Gibbs sampler in Mplus. In order to estimate the model parameters the following informative prior distributions were used (Dai et al., 2009; Li et al., 2009):

a_{jg} ~ Normal (0, 1) and $a_{jg} > 0, j = 1, \ldots, k$

b_{jg} ~ Normal (0, 1), $j = 1, \ldots, k$

θ_{ig} ~ Normal (0, 1), $i = 1, \ldots, n$

β_1 ~ Normal (0, 1),

where a_{jg} is the item discrimination for latent group g; b_{jg} is the item difficulty for latent group g; θ_{ig} is the person ability for latent group g; β_1 is the slope of the simple logistic regression and represents the loading of the latent variable on the covariate; k denotes the number of items; n denotes the number of examinees in each latent group g; g denotes latent groups and is set to 2 in the current study.

The MCMC estimates of the parameters were sampled from the posterior distribution after each iteration. To obtain the estimates of the parameters the means of the sampling iterations were calculated after discarding the burnt-in iterations.

3.3 EVALUATION OF OUTCOMES

The performance of the mixture 2PL IRT model on latent DIF detection was evaluated in terms of three types of outcome measures: 1) Accuracy of latent membership classification; 2) Estimation error at the scale level; 3) Latent DIF detection.

3.3.1 Recovery Analysis

The purpose of the recovery analysis is to determine how well the generating parameters are recovered based on the simulated data. Consistent with prior studies, the recovery of the latent class membership and the recovery of the simulated parameters were examined as a measure of the model effectiveness.

The recovery of the latent group structure was measured by the proportion of examinees assigned to the correct/simulated latent groups. This was conducted by comparing the estimated latent group membership with the simulated group membership.

The recovery of the simulated parameters was measured at the scale level, which was different from a direct measure of the magnitude differences between the parameter estimates and the generating parameters. Based on the parameter estimates and the generating parameters, the expected scores corresponding to each specific ability location were computed respectively. The general equation can be represented as:

$$E\left(y_j\middle|\theta\right) = \sum_{k=1}^{k_j} kp_{jk}(\theta), \tag{17}$$

where $E\left(y_j\middle|\theta\right)$ is the expected score for item *j*; *k* represents the scoring level for item *j*; p_{jk} is the probability of responding to item *j* at the scoring level *k*. Since items are dichotomously scored, *k* equals 2, and equation (19) is simplified to:

$$E\left(y_j \middle| \theta\right) = p(y_j = 1|\theta), \tag{18}$$

where p is the probability of an examinee answering the item correctly given his/her ability level θ. Since the model parameters were estimated in Mplus, the IRT parameterization was considered to transfer the Mplus output parameters into the item discrimination and difficulty parameters. By setting the factor mean to 0 (α = 0) and the factor variance to 1 (ψ = 1), the item discrimination and difficulty parameters were represented as:

$$a_{jg} = \lambda_{jg}, \tag{19}$$

$$b_{jg} = \frac{\tau_{jg}}{\lambda_{jg}}, \tag{20}$$

where a_{jg} is the discrimination parameter for item *j* in group g, b_{jg} is the difficulty parameter for item *j* in group g, λ_{jg} is the factor loading of item *j* in group g, and τ_{jg} is the threshold of item *j* in group g. λ_{jg} and τ_{jg} are read from the Mplus outputs. Therefore, the computational equation for the expected scores was represented as:

$$E\left(y_j \middle| \theta\right) = \frac{\exp[\lambda_{jg}\left(\theta_{ig} - \frac{\tau_{jg}}{\lambda_{jg}}\right)]}{1 + \exp[\lambda_{jg}\left(\theta_{ig} - \frac{\tau_{jg}}{\lambda_{jg}}\right)]}, \tag{21}$$

The expected scores were then summed across all the items to a total expected score y_t, which was used to quantify the differences between the generating parameters and their estimated values based on one of the most frequently used statistic – root mean squared error (RMSE):

$$RMSE(y_{estimated}, y_{true}) = \sqrt{\frac{\sum_{i=1}^{n}(y_{i,estimated} - y_{i,true})^2}{n}}, \tag{22}$$

where *n* denotes the number of subjects. RMSE can be considered as a measure of the absolute accuracy in parameter estimation. It was calculated for each replication of each condition, and then computed by taking the average value across the converged replications in each cell. Furthermore, RMSE can be computed across the simulation cells as a measure of the overall accuracy of the parameter estimation. Evaluating the accuracy of the parameter estimation at scale level through expected scores had two benefits. First, it overcomes the metric identification issues as was discussed in section 2.4. The calculation of the expected scores involves both the item and the ability parameters, which helps put these parameters on the same scale. Second, since the DIF magnitude is often assessed using the differences in the expected item scores between groups, accurate parameter estimation in terms of the expected scores ensures the appropriateness of further analysis of DIF at the scale level (i.e., DIF amplification or cancellation).

3.3.2 Latent DIF Detection

The main purpose of this measure is to examine how well DIF items can be identified using the mixture 2PL IRT model. Based on the idea of signal detection theory, four indices can be used to quantify the ability of the model in discerning among information-bearing patterns including: 1) number and percentage of DIF items that are identified as displaying DIF, known as power; 2) number and percentage of non-DIF items that are not identified as displaying DIF, known as correct non-DIF decision; 3) number and percentage of DIF items that are not identified as displaying DIF, known as type II error or false negative; 4) number and percentage of non-DIF items that are identified as displaying DIF, also known as type I error or false positive (Maij-de Meij, et al., 2010) (see Table 3.4). Last, the cell of “Power” and “Correct Non-DIF Decision” can be combined and indexed as “Correct Decision”. Given that false positive and false negative are considered relative to power and correct non-DIF decision, it is redundant to report all four indices. Thus, power, correct non-DIF decision, and correct decision are selected as the measures of latent DIF detection in the present study.

Table 3.4. Indices for DIF Detection

		Item	
		DIF	**Non-DIF**
Identification	**DIF**	Power	False Positive
	Non-DIF	False Negative	Correct Non-DIF Decision

3.4 SOLUTIONS TO LABEL SWITCHING

As was discussed in section 2.4, different types of strategies have been used to deal with IRT model identification issues. The person centering method was selected in the present study to deal with this issue. To be specific, the 70% group was selected as the reference latent group and the 30% group as the focal latent group. For the no impact condition, the ability distribution of both the reference and the focal groups were fixed to N(0, 1). For the impact condition, the ability of the focal group was fixed to N(0, 1), and the ability of the reference group was unconstraint. In this case, the first item, which was a non-DIF item, was used as an anchor item in Mplus by default.

Two types of label switching are often unavoidable during the Bayesian MCMC estimation, which are distinguished as the within-chain versus the between-chain label switching. The within-chain label switching is a more serious problem because the labels of the latent classes switch within a single analysis and thus distorts the interpretation of latent classes. The between-chain label switching means that the MCMC chain has converged to one peak in the posterior, while another chain has converged to another peak. This is commonly observed and would cause confusions in interpretation because the order of latent classes switches across chains.

By using Mplus for estimation, both the within-chain and between-chain label switching need to be addressed (Asparouhov & Muthén, 2008). To evaluate the within-chain label switching the posterior densities of the latent group membership parameters were monitored. If multiple modes are present in a Markov chain, it implies that the within-chain label switching occurs. In that case, the model is considered not a good fit to the data. The solution is to exclude these chains from the simulation replications, and calculate the percentage of the replications that contains these chains to see if they are a small number of the total simulation replications (Cho et al., 2013; Dai, 2009; Li et al., 2009; Li, 2015).

To handle the between-chain label switching, two strategies were adopted. First, the starting values were specified for the model parameters to help reduce the occurrence of this type of label switching. Second, class sizes were monitored throughout the analysis and model parameter estimates obtained from each replication were compared with the generating parameters for each latent group. The percentage of the replications that contained these chains was also evaluated to determine if they were a small number of the total simulation replications.

3.5 SIMULATION PROCESS

The first part of the simulation was to generate item parameters in terms of their specified distributions, and then generate item responses and obtain the expected scores for each examinee. Two sets of generated parameters are presented in Appendix Table A.3 and Table A.4, which provided an example of the manipulation of DIF size, number of DIF items, and DIF type. The DIF items are shown at the bottom of the Table A.3 and Table A.4. The second part was to fit the simulated response data using the mixture 2PL IRT model with and without the dichotomous covariate by implementing two estimation methods. The last part was the analysis of DIF in terms of the assigned latent group membership.

To be specific, the model parameters and response data generation process included the following steps:

1) The a and b parameters were generated for each latent group with the sample size of 1400 for the reference group and 600 for the focal group. The *a* parameter followed the distribution of lognormal (0, 0.5), and the b parameter followed the distribution of uniform (-2, 2);
2) 6 or 12 items were assigned to be the DIF items, corresponding to the DIF item proportions of 15% and 30%. For the uniform DIF condition, the *a* parameter was kept the same for both latent groups. The *b* parameter of the DIF items for the focal latent group g_2 was calculated as $b_2 = b_1 + \Delta b$, where b_1 was the b parameter for the reference group, and $\Delta b = 0.5$ or 1. The values of the *b* parameter for the rest of the items were kept the same between the two latent groups, $b_2 = b_1$. Similarly, for the non-uniform DIF condition, the *b* parameter was kept the same for both latent groups. The *a* parameter of the DIF items for the focal latent group g_2 was calculated as $a_2 = a_1 - \Delta a$, where a_1 was the a parameter for the reference group, and $\Delta a = 0.5$ or 1. The values of the *a* parameter for the rest of the items were kept the same between the two latent groups: $a_2 = a_1$.
3) For the "no impact" condition, the ability distributions of both latent groups (g_1 and g_2) were the same: $\theta_1 \sim N(0, 1)$ and $\theta_2 \sim N(0, 1)$. For the "impact" condition, the ability distributions were: $\theta_1 \sim N(1, 1)$ for the reference group and $\theta_2 \sim N(0, 1)$ for the focal group.

4) Simple logistic regression was used to model the relationship between the covariate and the latent group membership: $\text{logit}(\pi) = \beta_0 + \beta_1 D$. The covariate was included in the generation of response data, and the generation of the relationship between the covariate and the latent variable was completed by specifying the number of examinees in each cell of Table 3.3.
5) The probability of a correct response to each test item was calculated based on the 2PL IRT model using the generated *a* and *b* parameters, and the item responses were generated given the values of these probabilities.
6) Referring to Equation (23), the expected score of each item for each examinee was computed. The expected scores were then summed across test items.

The second part of the simulation was to estimate the data generating model. The mixture 2PL IRT model was used to estimate the item and ability parameters. The estimated expected scores were computed based on the estimated parameters. As was discussed in section 2.5.2, the one-step approach was adopted because the focus of this study is on the covariate effect on parameter estimation and classification rather than the mechanism/cause of DIF. The covariate was directly integrated into the mixture model, contributing to the estimation of latent class membership and model parameters (Smit et al., 1999; Dai, 2009; Li et al., 2015). Ability estimates were obtained by requesting "SAVE=FSCORES" in the SAVEDATA command in Mplus. The estimated item parameters were the factor loading λ and threshold τ. The IRT parameterization was used to transfer factor loadings and thresholds into the IRT *a* and *b* parameters using Equation (21) and (22) (Asparouhov & Muthén, 2016). The estimation method was specified through "ESTIMATOR = MLR" and "ESTIMATOR = BYES" in Mplus, respectively.

The last part of the simulation was to evaluate the recovery of the latent structure and model parameters as well as the detection of DIF referring to the outcome measures. The DIF detection was performed given the estimated latent class membership. This procedure was performed in R with the difLogistic function from the difR package.

3.6 VALIDATION OF SIMULATION PARAMETERS

To ensure that item parameters were appropriately generated, a validation procedure for one of the simulation conditions was conducted to evaluate the recovery of the simulation parameters. The simulation condition selected was when the DIF size was .5, the proportion of DIF items was 15%, the DIF items were uniform, the odds ratio was 2, there was no group impact, no covariate was included, and the MLE was used. Given the generated group membership, the 2PL IRT model was set up in Mplus to estimate the item parameters for each of the latent groups separately. In this analysis, the item parameters for the non-DIF items were constrained to be equal between the two groups, while the items parameters for the DIF items remained unconstrained. These parameter estimates were compared with the simulation parameters for each group, and their differences were quantified using RMSE as a measure of parameter simulating errors. The parameter validation results for each latent group were summarized in Appendix Table A.1. For the reference latent group, the RMSE was .085 for the discrimination parameter and was .106 for the difficulty parameter. For the focal latent group, the RMSE was .107 for the discrimination parameter and was .064 for the difficulty parameter.

3.7 COMPUTER PROGRAMS

Three computer programs were used in the current study. Model parameters and response data were generated in SAS 9.4. Parameter estimation based on the mixture 2PL model was performed in Mplus 7.4. Two estimation methods including MML and Bayesian estimation were fitted by specifying estimator = MLR and estimator = Bayes respectively (Muthén, 2010). The analysis of DIF was conducted in R by using the difLogistic function in the difR package, which is a popular method of detecting both the uniform and nonuniform DIF for dichotomously scored items. The sample codes for response data generated in SAS are presented in Appendix B. The sample codes for model estimation in Mplus with MLE and Bayesian estimation are presented in Appendix C and Appendix D.

4.0 RESULTS

The primary purposes of the present study is to compare the detection of uniform and non-uniform DIF between the mixture 2PL model without a covariate and the model with a covariate. The secondary purpose is to examine how the detection of latent DIF is affected by the parameter estimation method employed. As was described in Chapter 3, seven factors were manipulated, including the strength of the relationship between the covariate and the latent group membership, the type of DIF, the proportion of items with DIF, the magnitude of DIF, the group impact, the model specification, and the estimation method. The performance of the mixture 2PL model on latent DIF detection is evaluated in terms of three categories of outcome measures, including the accuracy of latent membership classification, the model parameter recovery at the scale level, and the detection of DIF. The results of these outcome measures are summarized below.

4.1 ANALYTICAL PLAN

The model diagnostic was conducted to evaluate the convergence and the label switching issues. A single simulation condition was selected for this purpose, which was used to evaluate the occurrence of label switching in other simulation conditions. The non-converged replications were excluded from the final estimation of the model parameters. The model diagnostic is described in section 4.2.

The manipulated factors can be categorized into the within-replication factors and the between-replication factors. Two within-replication factors were the analyzing model and the estimation method. Within each replication each generated data set was analyzed using the mixture 2PL model with and without the covariate, and each model was estimated using MLE and Bayes estimator respectively. Five between-replication factors included the type of DIF, the proportion of items with DIF, the group impact, the strength of the relationship between the covariate and the latent group membership, and the magnitude of DIF. The abbreviations and coding of each factor are listed in Table 4.1:

Table 4.1. Variable Abbreviations and Coding

Manipulated factors	**Abbreviations**	**Coded as "0"**	**Coded as "1"**
Within-replication factors:			
Model specification	Model	NoCov	Cov
Estimation Methods	Estimation	MLE	Bayesian
Between-replication factors:			
Type of DIF	DIFtype	Uniform	Non-uniform
Proportion of DIF items	DIFnum	15%	30%
DIF magnitude	DIF	0.5	1
Group impact	Impact	N(0, 1), N(0, 1)	N(1, 1), N(0, 1)
Magnitude of odds ratio	OR	2	8

Given the mixed factorial design of the present study, the mixed analysis of variance (mixed ANOVA) was used to analyze data in SPSS 24.0. This analysis was repeated for each of the outcome measures. The assumptions of the mixed ANOVA include the normality of the dependent variables, homogeneity of the covariance matrices, sphericity, and outliers. Since the grouping variables have only two levels, the Mauchly's test of sphericity was not performed. The rest of the assumptions were examined through Shapiro-Wilk test of normality, Box's M test of homogeneity of covariance matrices, Levene’s test for homogeneity of variances, and the SPSS Explore procedure.

The focus of the analysis was the two-way and three-way interaction effects among the within-replication and the between-replication factors. Given the complexity of interpretation and negligible effect sizes, the four-way or more-way interactions were not considered. The significant three-way interaction effects were followed by simple two-way interaction analyses to examine the two-way interaction effects at each level of the third factor. Post hoc analysis was not needed because the manipulated factors only have two levels. The effect sizes were judged in terms of the partial eta squared statistic:

$$\eta^2=\frac{SS_{effect}}{SS_{effect}+SS_{error}}, \tag{23}$$

The suggested norms for the partial eta squared are: small = .01, medium = .06, large = .14 (Cohen, 1988). In the consideration of empirical significance, the significant effects were reported only when the partial $\eta^2 \geq .06$.

4.2 EVALUATION OF CONVERGENCE AND LABEL SWITCHING

The convergence issue was examined for the MLE and Bayesian estimation respectively. It was found that for the MLE no convergence issues were observed in any replications. For the Bayesian estimation, a total of 20,000 iterations were specified in Mplus. Half of the iterations were discarded by default in Mplus as burn-in, and 10,000 post burn-in values were used to obtain the final parameter estimates (5,000 iterations in each chain). Thinning of the posterior draws was set to 50 (Finch & French, 2012), which means every 50th iteration was recorded for estimation process. The use of the "thin" option was to reduce the computer storage. Each simulation replication took about 15 minutes to run in Mplus.

Given the large number of simulation conditions, only one conditions was selected to make an evaluation of the label switching issues in Bayesian estimation. This condition was chosen by setting the proportion of DIF items = 15%, DIF type = uniform, DIF size = .5, OR = 2, and group impact = N(0, 1) & N(0, 1). The combination of these factor levels formed a condition that was most likely to suffer from label issues. The posterior density plots of the group membership and the Mplus trace plots were inspected, and the parameter estimates were compared with the simulation parameters.

In the cased of two chain converged, the results included the within-chain label switching, collapsed chains, poor mixing chains, and recovered chains. In the case of two chain not converged, the results included both the within- and between- chain label switching, collapsed chains, poor mixed chains, and recovered chains. In addition to the label switching chains and recovered chains, the collapsed chains occurred when the two chains exhibited convergence to a stationary distribution and produced a solution with essentially one latent class. In this case, the proportion of examinees classified into one group was almost 100% and was close to 0% for the other group. The occurrence of the collapsed chains implied that information was insufficient to separate examinees into two latent classes. The poor mixing chains occurred when the two chains fluctuated non-systematically within a wide range of possible values for the parameter.

The percentages of the occurrence of each type of results were summarized in Appendix Table A.5. As shown, approximately 73% of the chains recovered the latent structure appropriately when averaged across the convergent and non-convergent MCMC chains. The occurrence of the within-chain label switching was as low as 1.027%. The assigning of starting values to model parameters was to help reduce the occurrence of between-chain label switching, which led to approximately 11% of the chains not converged. Approximately 14% of the chains were collapsed into a single class solution, while the percentage of poor mixing chains was as low as approximately 2%.

Given the total number of replications was as large as 6400, the replications that contained non-convergent chains were excluded from the analysis, which was 19% of the 50 replications in this condition. In terms of previous research, the relabeling of the groups to match with the simulating group membership did not solve the between-chain label switching issue. In Dai's study (2009), only 21% of the between-chain label switching runs out of 720 replications were resolved by adding additional group membership information. In addition, the condition selected for the model diagnostic represented the situation where the label switching issue was most likely to occur. It is expected that for other simulation conditions the occurrence of the between-chain label switching would be less. The final analysis of all the simulation conditions found that the number of replications used for parameter estimation ranged from 41 to 48 with 45 replications were kept for most of the conditions.

4.3 RECOVERY OF LATENT GROUP MEMBERSHIP

The recovery of the latent group membership was evaluated through the accuracy of latent group classification. It was defined as the proportion of examinees correctly assigned to the simulated latent groups. The descriptive statistics of the latent group classification accuracy are presented in Table 4.2. As shown, the classification rates were moderately high, ranging approximately from

.600 to .800. The latent group assignment was more accurate when the covariate was included or when the Bayes estimator was used. The accuracy of classification tended to be higher when the covariate was more related to the latent group membership. In addition, the recovery of the latent structure was better when there was a separation between the two latent groups or when the DIF magnitude was larger. It was also noted that the recovery of the latent group membership was better for the uniform DIF items compared to the non-uniform DIF items. The average classification accuracy for each of the simulation conditions is presented in Table A.6 in Appendix A.

Table 4.2. Descriptive Statistics of Latent Group Classification Accuracy by Factors

Manipulated factors	**Levels**	**M**	**SD**
Within-replication factors:			
Model specification	NoCov	.673	.142
	Cov	.719	.133
Estimation Methods	MLE	.682	.149
	Bayesian	.725	.123
Between-replication factors:			
Type of DIF	Uniform	.713	.141
	Non-uniform	.678	.136
Proportion of DIF items	15%	.656	.132
	30%	.736	.135
DIF magnitude	0.5	.610	.128
	1	.782	.088
Group impact	N(0, 1), N(0, 1)	.673	.139

	N(1, 1), N(0, 1)	.729	.136
Magnitude of odds ratio	2	.685	.137
	8	.708	.141

It has been shown that ANOVA is robust to moderate violation of normality (Glass, Peckham, & Sanders, 1972). Therefore, transformations were not performed to correct for the violation of normality. The other assumptions were met, and potential outliers were excluded from the analyses. Table 4.3 summarizes the mixed ANOVA results.

Table 4.3. Main and Interaction Effects of Factors on Classification Accuracy

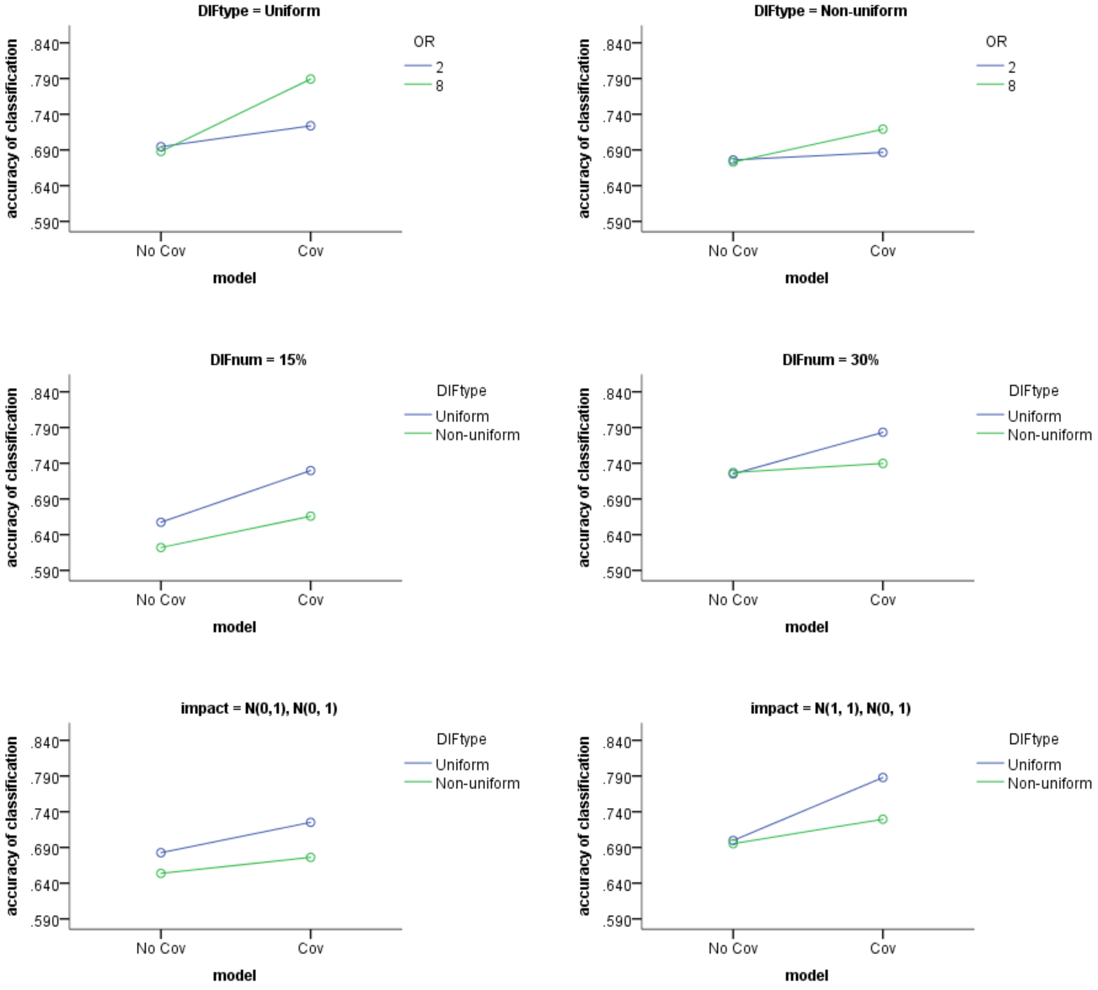
Source	**F value**	**p value**	**Partial $\boldsymbol{\eta}^2$**
Within-replication factors:			
Model	203.717	<.001	.322
Model × OR	67.826	<.001	.137
Model × DIFtype	31.957	<.001	.069
Model × OR × DIFtype	42.427	<.001	.101
Model × DIFtype × DIFnum	36.282	<.001	.079
Model × DIFtype × Impact	34.223	<.001	.072
Estimation	55.706	<.001	.115
Estimation × DIFtype	30.402	<.001	.066
Estimation × DIFtype × Impact	38.164	<.001	.087
Between-replication factors:			
OR	35.812	<.001	.077
DIFtype	90.893	<.001	.175

DIFnum	412.509	<.001	.491
DIF	1914.330	<.001	.817
Impact	140.103	<.001	.247
DIRnum × DIF	87.053	<.001	.169
DIFnum × DIFtype × DIF	60.128	<.001	.103

The two-way interactions of the model and the relationship between the covariate and the latent variable, the model and the DIF type, as well as the estimation method and the DIF type were all dependent on another factor. To be specific, the interaction of the model and the relationship between the covariate and the latent group membership was shown to be dependent on the DIF type. The simple two-way interaction analysis showed that this interaction was significant for both the uniform and the non-uniform DIF items with larger effect size for the uniform DIF items, F = 49.228, p < .001, partial η^2 = .141, F = 13.142, p < .001, partial η^2 = .028. In other words, for both types of DIF items the improvement of the classification accuracy by including the covariate into the model was significantly larger when this covariate was more related to the latent group membership.

The interaction of the model and the DIF type depended on the number of the DIF items, which had larger effect size when more DIF items were included, F = 18.322, p < .001, partial η^2 = .059, F = 37.657, p < .001, partial η^2 = .109. This indicated that the improvement of the classification accuracy by including the covariate into the model was significantly better for the uniform DIF items particularly when there were more DIF items. The interaction of the model and the DIF type also depended on the difference in group abilities, F = 15.078 p < .001, partial η^2 = .046, F = 22.485, p < .001, partial η^2 = .078. Regardless of the group impact, the improvement of the classification accuracy by including the covariate was significantly better for the uniform DIF items.

The interaction of the estimation method and the DIF type depended on the difference in group abilities. The simple two-way interaction analysis indicated that this interaction was significant for both the no impact and the impact groups, F = 12.789, p < .001, η^2 = .053, F = 7.360, p = .007, η^2 = .031. Regardless of the group impact, the improvement of the classification accuracy by using the Bayes estimator was significantly better for the uniform DIF. The graphic representations of the three-way interactions among the within- and the between-replication factors are shown in Figure 4.1.



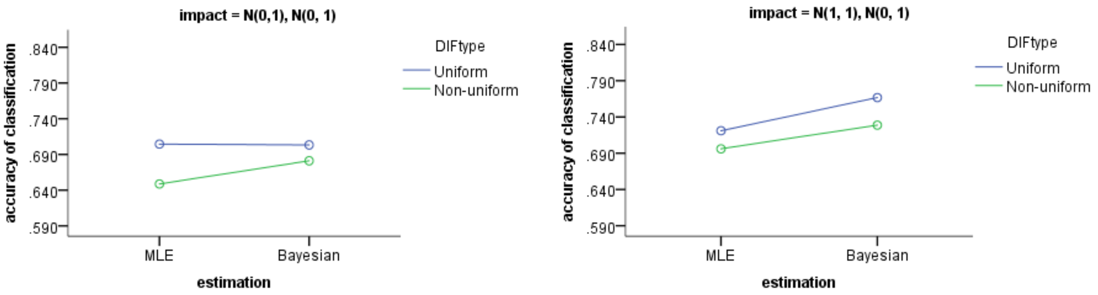


Figure 4.1. 3-Way Interactions of within-Replication Factors: Classification Accuracy

The results of the effects for the between-replication factors showed that the interaction of the DIF type and the proportion of DIF items depended on the DIF magnitude. The simple two-way interaction analysis indicated that this interaction was significant for both levels of the DIF magnitudes, F = 25.390, p < .001, η^2 = .024, F = 64.324, p < .001, η^2 = .063. This implied that the improvement of the classification accuracy by using more DIF items depended on the nature of the DIF items. This three-way interaction is presented in Figure 4.2.

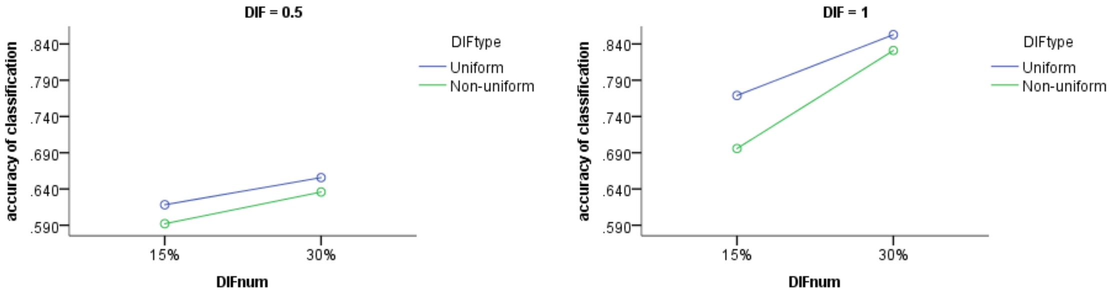


Figure 4.2. 3-Way Interactions of between-Replication Factors: Classification Accuracy

4.4 RECOVERY OF MODEL PARAMETERS

The joint recovery of the model parameters was evaluated in terms of RMSE. It was calculated by quantifying the differences between the expected scores based on the estimated parameters and the expected scores based on the simulation parameters. The descriptive statistics of RMSE averaged across the simulation cells are presented in Table 4.4. As shown, the values of RMSE ranged from .250 to .290. Averaging across the simulation conditions, the recovery of the model parameters was better (lower RMSE value) when the parameters were estimated using the 2PL model with the covariate and the Bayes estimator. The absolute errors in parameter estimation were relatively small when the DIF size was large or more DIF items were included in the test. Likewise, the recovery of the model parameters tended to be better for the uniform DIF items or when the two groups differentiated in their abilities. The average RMSE values for each of the simulation conditions are presented in Table A.7 in Appendix A.

Table 4.4. Descriptive Statistics of RMSE by Factors

		RMSE	
Manipulated factors	**Levels**	**M**	**SD**
Within-replication factors:			
Model specification	NoCov	.282	.107
	Cov	.263	.038
Estimation Methods	MLE	.288	.105
	Bayesian	.252	.009
Between-replication factors:			
Type of DIF	Uniform	.271	.068

	Non-uniform	.275	.092
Proportion of DIF items	15%	.281	.106
	30%	.265	.043
DIF magnitude	0.5	.277	.068
	1	.269	.093
Group impact	N(0, 1), N(0, 1)	.282	.059
	N(1, 1), N(0, 1)	.263	.098
Magnitude of odds ratio	2	.273	.078
	8	.272	.085

Table 4.5 summarizes the significant main and interaction effects for the manipulation factors on RMSE. As shown, large effect sizes were achieved for most of the interactions. Simple analyses were performed as the follow-up to the significant three-way interactions. For both of the uniform and non-uniform DIF items, the reduction in the errors of the parameter estimation by including a covariate into the model significantly depended on the relationship between this covariate and the latent group membership, F = 8.731, p = .003, η^2 = .027, F = 6.196, p = .014, η^2 = .025. This pattern was also observed for both levels of the DIF item proportions, F = 5.887, p = .016, η^2 = .018, F = 5.147, p = .015, η^2 = .012. The interactions between the model and the DIF size were also significant for both the no impact and the impact groups, F = 6.242, p = .013, η^2 = .019, F = 59.538, p < .001, η^2 = .200. The interactions between the model and the DIF type were significant regardless of the number of DIF items, F = 20.834, p < .001, η^2 = .081, F = 17.234, p < .001, η^2 = .064.

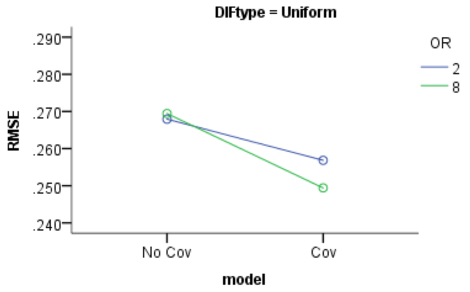
The reduction of the errors in parameter estimation by using the Bayes estimator was significantly more pronounced for the uniform DIF items compared to the non-uniform DIF items, which was found regardless of the covariate-membership association, F = 19.978, p < .001, η^2 = .077, F = 9.589, p = .002, η^2 = .039, or the group impact, F = 7.388, p = .007, η^2 = .030, F = 60.695, p < .001, η^2 = .203. The interaction of the estimation method and the DIF size was significant for the impact group only, F = 71.301, p < .001, η^2 = .231.

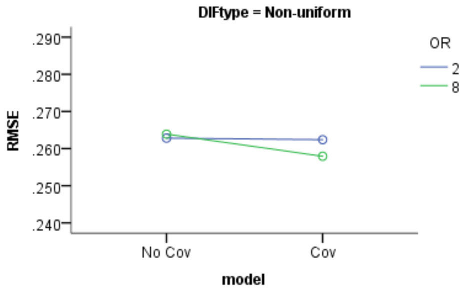
Referring to RMSE, the interaction of the two within-replication factors was found to be significant relative to the DIF size and the group impact. The reduction in estimating errors when including the covariate was more pronounced when using the MLE, which was observed for both of the uniform and the non-uniform DIF items, F = 146.901, p < .001, η^2 = .381, F = 8.152, p = .005, η^2 = .033, and for both of the no impact and the impact groups, F = 20.636, p < .001, η^2 = .081, F = 42.911, p < .001, η^2 = .152. The graphic representations of the three-way interactions among the within- and the between-replication factors are shown in Figure 4.3.

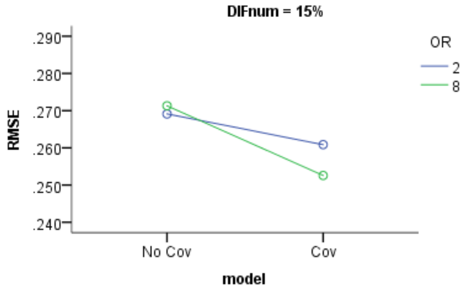
Table 4.5. Main and Interaction Effects of Factors on RMSE

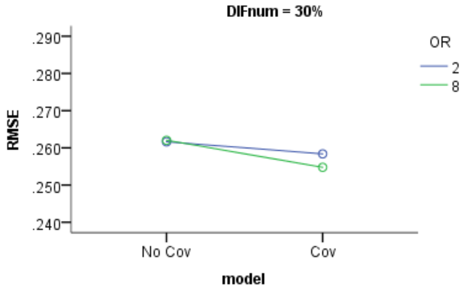
Source	**F value**	**p value**	**Partial $\boldsymbol{\eta^2}$**
Within-replication factors:			
Model	77.650	<.001	.148
Model × OR	30.874	<.001	.064
Model × DIFtype	81.387	<.001	.157
Model × DIFnum	70.101	<.001	.125
Model × DIF	57.173	<.001	.113
Model × Impact	50.212	<.001	.101
Model × OR × DIFtype	27.194	<.001	.061

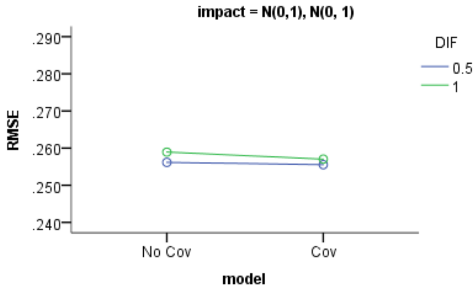
Model × OR × DIFnum	31.357	<.001	.065
Model × DIF × Impact	68.425	<.001	.132
Model × DIFtype × DIFnum	57.223	<.001	.118
Model × Impact × OR	70.895	<.001	.137
Estimation	85.157	<.001	.163
Estimation × DIFtype	107.560	<.001	.194
Estimation × DIFnum	215.284	<.001	.325
Estimation × DIF	63.477	<.001	.124
Estimation × Impact	77.771	<.001	.148
Estimation × DIFtype × OR	29.832	<.001	.065
Estimation × DIFtype × Impact	128.579	<.001	.223
Estimation × DIF × Impact	72.535	<.001	.139
Model × Estimation × DIFtype	135.543	<.001	.232
Model × Estimation × Impact	102.000	<.001	.185
Between-replication factors:			
DIFnum	35.489	<.001	.073
DIF	152.368	<.001	.254
Impact	153.312	<.001	.255
DIFtype × Impact	40.738	<.001	.083
DIFnum × Impact	48.537	<.001	.098
DIFnum × DIFtype × DIF	99.555	<.001	.182

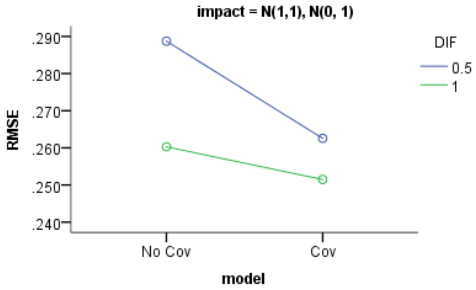

DIFtype = Uniform
RMSE
.290
.280
.270
.260
.250
.240
No Cov
Cov
model
OR
2
8

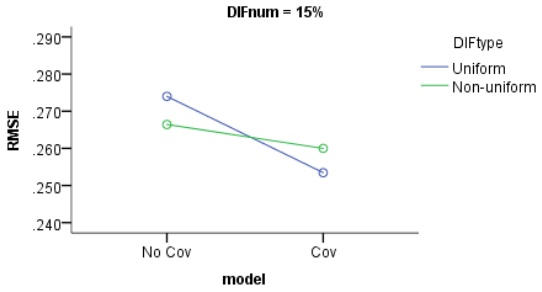

DIFtype = Non-uniform
RMSE
.290
.280
.270
.260
.250
.240
No Cov
Cov
model
OR
2
8

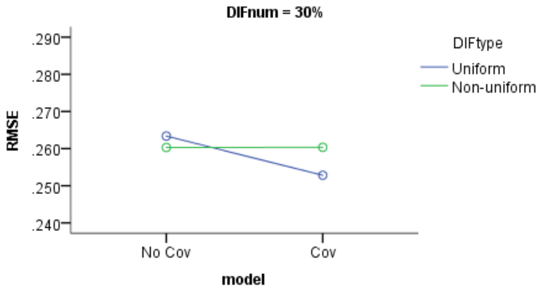

DIFnum = 15%
RMSE
.290
.280
.270
.260
.250
.240
No Cov
Cov
model
OR
2
8

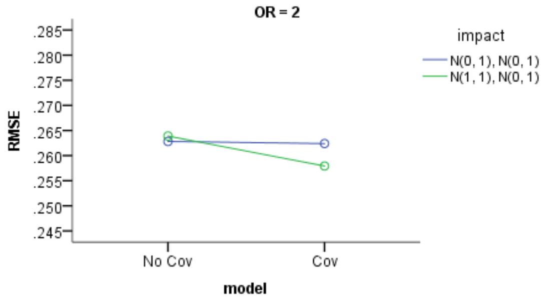

DIFnum = 30%
RMSE
.290
.280
.270
.260
.250
.240
No Cov
Cov
model
OR
2
8

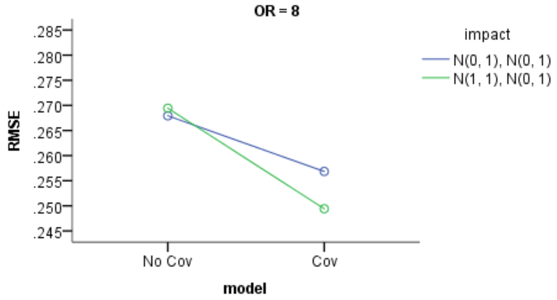

impact = N(0,1), N(0, 1)
RMSE
.290
.280
.270
.260
.250
.240
No Cov
Cov
model
DIF
0.5
1

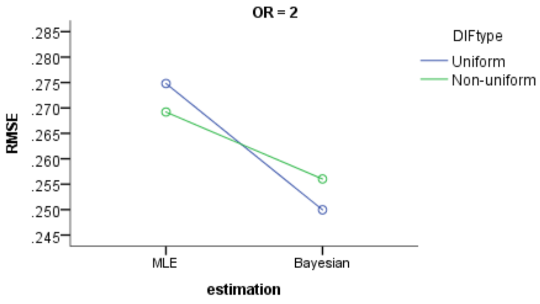

impact = N(1,1), N(0, 1)
RMSE
.290
.280
.270
.260
.250
.240
No Cov
Cov
model
DIF
0.5
1

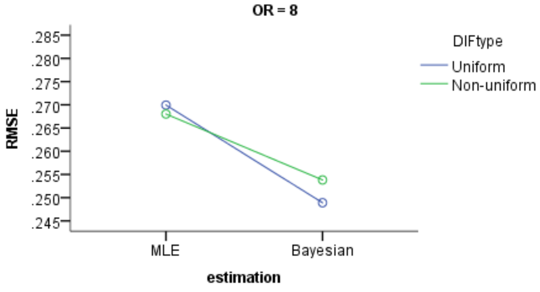

DIFnum = 15%
RMSE
.290
.280
.270
.260
.250
.240
No Cov
Cov
model
DIFtype
Uniform
Non-uniform

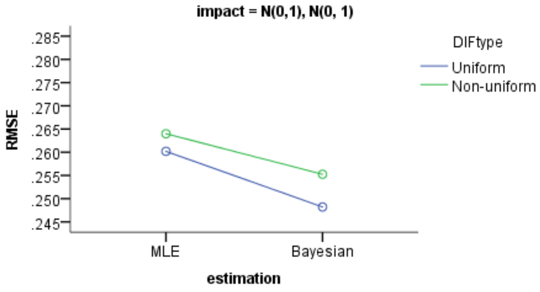

DIFnum = 30%
RMSE
.290
.280
.270
.260
.250
.240
No Cov
Cov
model
DIFtype
Uniform
Non-uniform

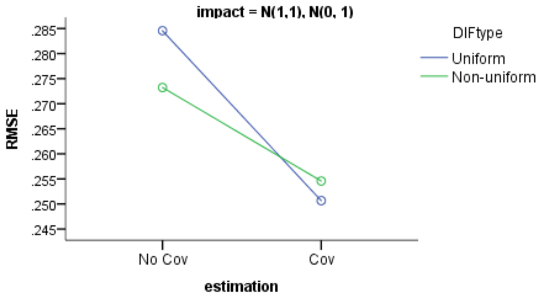

OR = 2
RMSE
.285
.280
.275
.270
.265
.260
.255
.250
.245
No Cov
Cov
model
impact
N(0, 1), N(0, 1)
N(1, 1), N(0, 1)

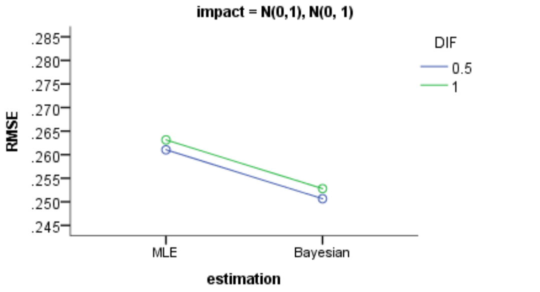

OR = 8
RMSE
.285
.280
.275
.270
.265
.260
.255
.250
.245
No Cov
Cov
model
impact
N(0, 1), N(0, 1)
N(1, 1), N(0, 1)

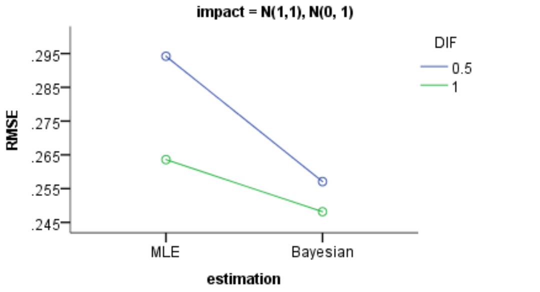

OR = 2

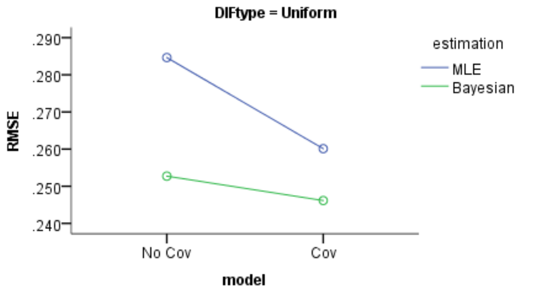

OR = 8

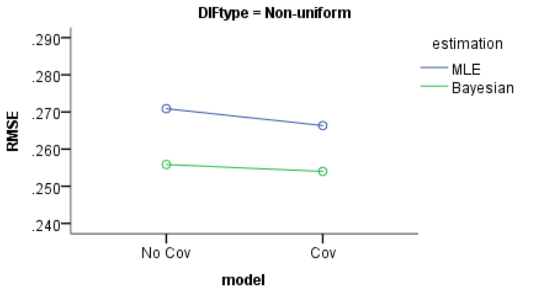

impact = N(0,1), N(0, 1)


impact = N(1,1), N(0, 1)


impact = N(0,1), N(0, 1)


impact = N(1,1), N(0, 1)


DIFtype = Uniform


DIFtype = Non-uniform

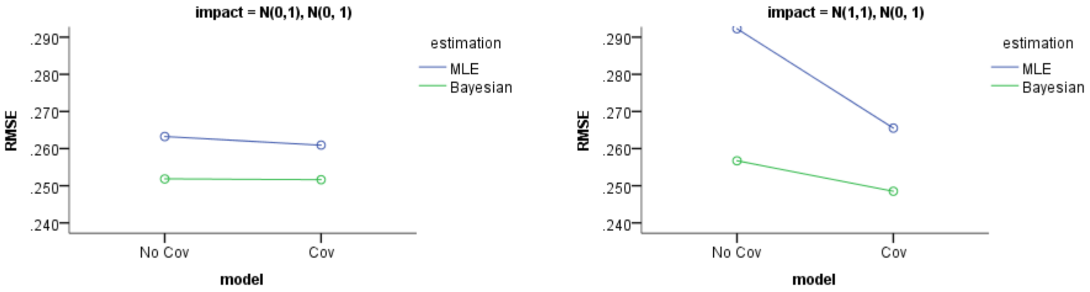


Figure 4.3. 3-Way Interactions of within-Replication Factors: RMSE

The three-way interactions of the between-replication factors are presented in Figure 4.4. The interactions between the covariate-membership relationship and the DIF size were significant for both the no impact and the impact groups. The interactions of DIF size and the proportion of DIF items were significant regardless of the DIF type or group impact.

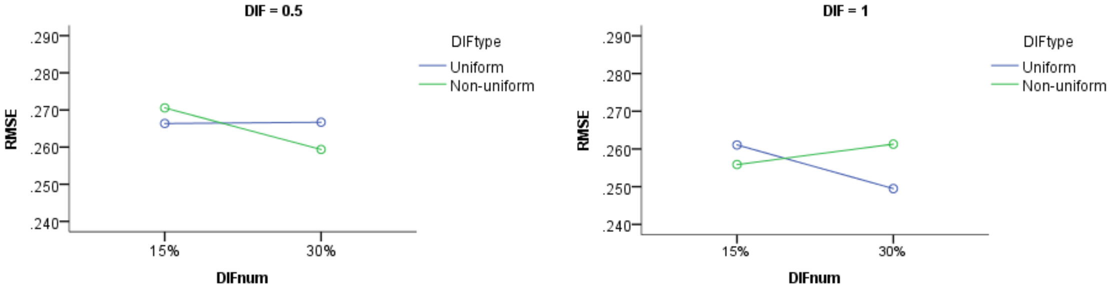


Figure 4.4. 3-Way Interactions of between-Replication Factors: RMSE

4.5 DETECTION OF LATENT DIF

The primary interest of the present study is to evaluate how well DIF items can be identified based on the latent grouping estimated by the mixture 2PL IRT model. Three indices are reported as a measure of the model's efficiency in DIF detection: power, correct non-DIF decision (non-DIF

CD), and correct decision (CD) rates. The descriptive statistics of these three indices for each within- and between-replication factor averaged across simulation conditions are presented in Table 4.6. These descriptive statistics provides the overall effect of each factor. As shown, all three indices showed the same direction of the group differences for each of the factors. Overall, the larger power, higher correct non-DIF decision, and higher correct decision rates were found for the conditions of mixture 2PL model with the covariate, Bayesian estimation, uniform DIF items, large number of DIF items, and large DIF size. In addition, the separation of the latent groups and the association between the covariate and the group membership tended to affect the detection of DIF. The analyses of interaction effects below disclosed how the effect of each factor was manifested relative to other factors.

The descriptive statistics of the power, correct non-DIF decision, and correct decision rates averaged across replications for each of the simulation conditions are presented in Table A.8, Table A.9, and Table A.10 in Appendix A. These descriptive statistics provides the interaction effects.

Table 4.6. Descriptive Statistics of DIF Detection Indices by Factors

		Power		**Non-DIF CD**		**CD**	
Manipulated factors	**Levels**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Within-replication factors:							
Model specification	NoCov	.645	.087	.857	.077	.089	.064
	Cov	.782	.062	.868	.074	.074	.061
Estimation Methods	MLE	.681	.095	.840	.071	.827	.055
	Bayesian	.804	.103	.856	.035	.857	.049
Between-replication factors:							
Type of DIF	Uniform	.723	.093	.867	.031	.857	.065

	Non-uniform	.683	.130	.855	.087	.842	.068
Proportion of DIF items	15%	.693	.090	.851	.109	.840	.049
	30%	.709	.111	.869	.112	.856	.077
DIF magnitude	0.5	.684	.127	.848	.108	.833	.078
	1	.761	.096	.851	.105	.839	.069
Group impact	N(0, 1), N(0, 1)	.687	.120	.837	.094	.828	.101
	N(1, 1), N(0, 1)	.749	.106	.869	.102	.858	.066
Magnitude of odds ratio	2	.677	.118	.831	.127	.827	.068
	8	.725	.108	.874	.078	.861	.059

A good number of significant interaction effects were found to be consistent among the three indices of the DIF detection. The mixed ANOVA results for each of the three indices are summarized in Table 4.7, Table 4.8, and Table 4.9.

4.5.1 Power

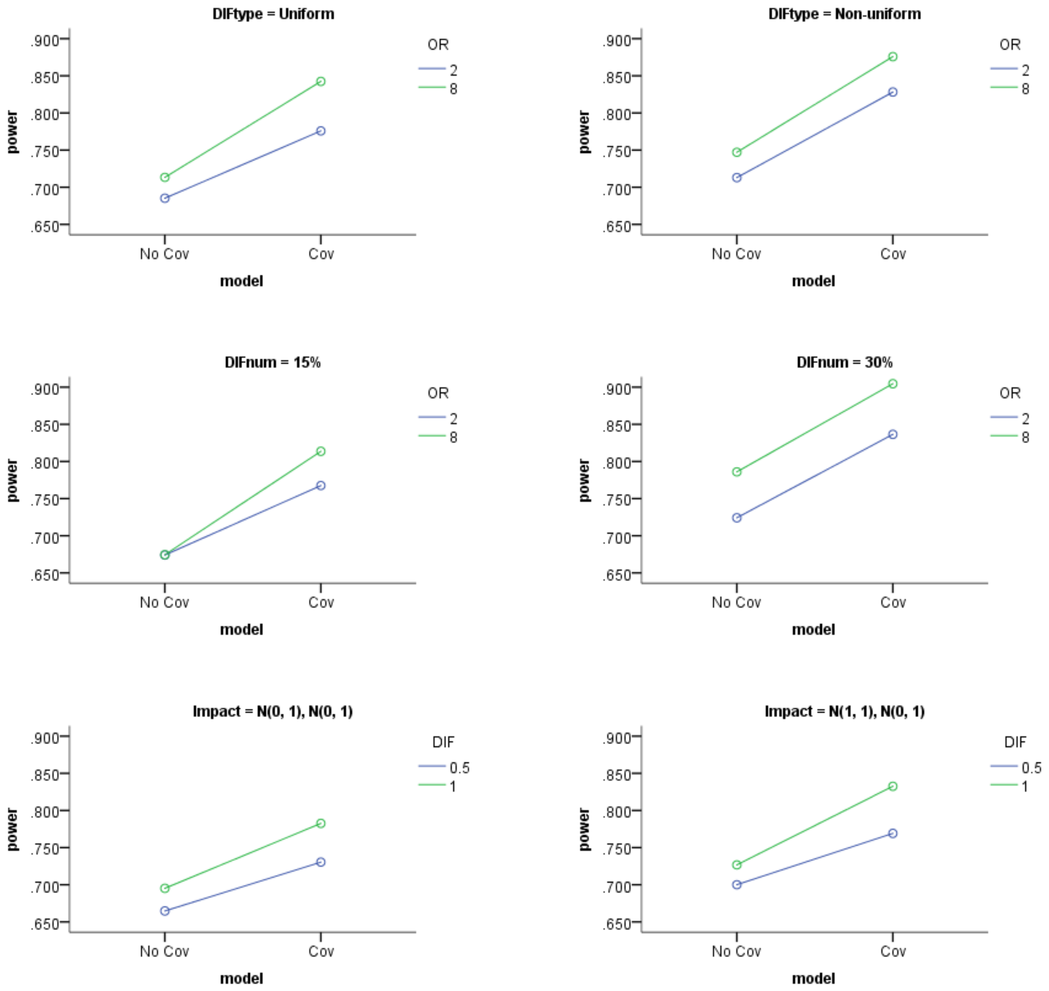
Table 4.7 shows the significant main and interaction effects with the partial η^2 larger than .60. The within-replication factor "model" had significant interactions with three between-replication factors including the relationship between the covariate and the latent group membership, the DIF type, and the proportion of DIF items. These two-way interactions depended on the third factor. To be specific, the interaction of the model and the covariate-membership relationship was significant when the DIF items were uniform or when there was a small number of DIF items, F = 17.132, p < .001, partial η^2 = .072, F = 25.145, p < .001, partial η^2 = .102. The interaction of the model and DIF size was significant regardless of the group differences with larger effect size when the group difference existed, F = 15.243, p < .001, partial η^2 = .046, F = 17.921, p < .001, partial η^2 = .078.

The interaction of the estimation method and the DIF type was significant only when the covariate was weakly related to the latent group membership, F = 28.267, p < .001, partial η^2 = .121. Similar as above, the interaction of the estimation method and the DIF type was significant regardless of the group differences with larger effect size when the group difference existed, F = 16.147, p < .001, partial η^2 = .064, F = 18.352, p < .001, partial η^2 = .083. No interaction was found between the two within-replication factors. The visual representations of these three-way interactions are shown in Figure 4.5.

Table 4.7. Main and Interaction Effects of Factors on Power

Source	**F value**	**p value**	**Partial η^2**
Within-replication factors:			
Model	81.292	<.001	.173
Model × OR	57.779	<.001	.128
Model × DIFtype	35.777	<.001	.074
Model × DIFnum	36.953	<.001	.075
Model × OR × DIFtype	26.257	<.001	.061
Model × OR × DIFnum	41.068	<.001	.092
Model × DIF × Impact	29.452	<.001	.066
Estimation	63.747	<.001	.164
Estimation × DIFtype	56.586	<.001	.126
Estimation × DIFtype × OR	46.951	<.001	.106
Estimation × DIFtype × Impact	40.298	<.001	.089

Between-replication factors:			
DIFtype	53.746	<.001	.123
DIFnum	29.314	<.001	.068
DIF	43.823	<.001	.097
DIFnum × DIFtype	58.328	<.001	.137
DIFnum × DIFtype × DIF	42.101	<.001	.094



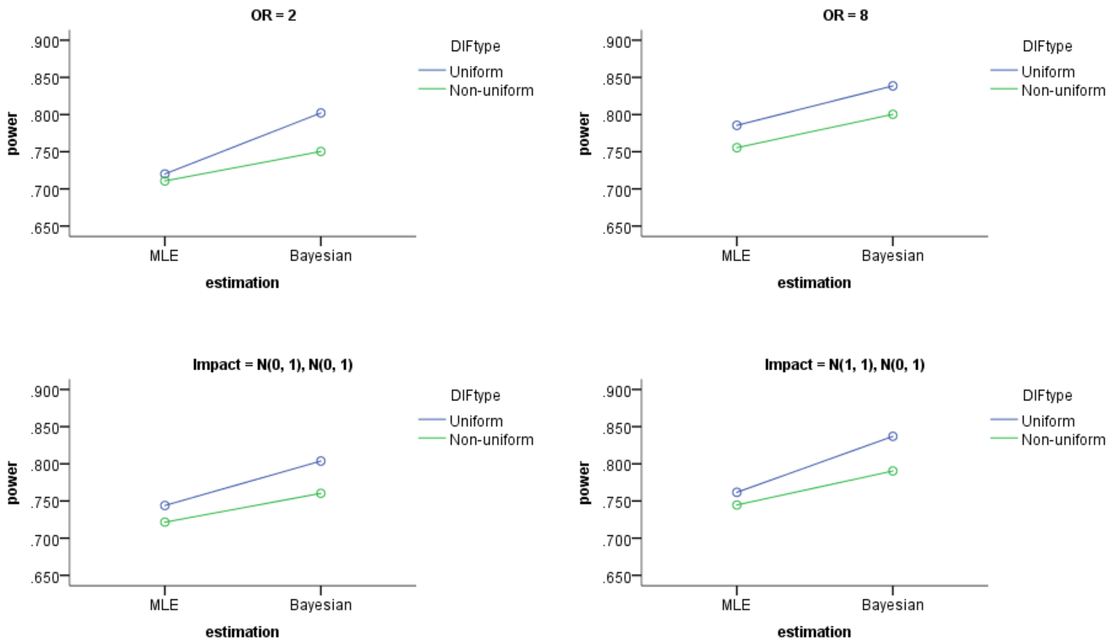


Figure 4.5. 3-Way Interactions of within-Replication Factors: Power

The results for the between-replication factors showed that the interactions of the DIF number and the DIF type were dependent on the DIF magnitude with larger effect size when the DIF magnitude was small, F = 73.025, p < .001, η^2 = .113, F = 32.315, p < .001, η^2 = .055. This three-way interaction is presented in Figure 4.6.

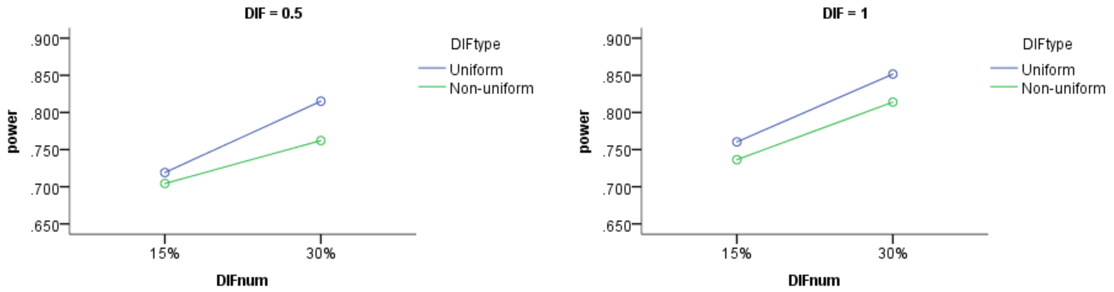


Figure 4.6. 3-Way Interactions of between-Replication Factors: Power

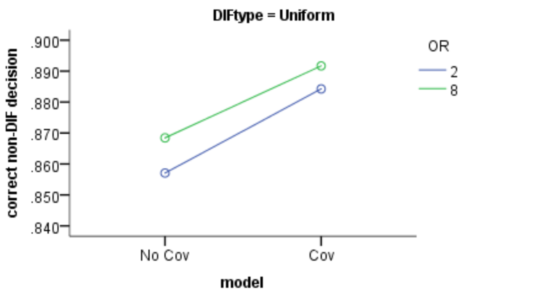
4.5.2 Correct Non-DIF Decision

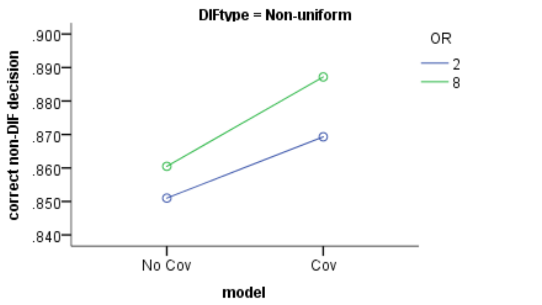
The results of the interaction effects for correct non-DIF decision were generally consistent with the results for power. The interactions of the within-replication factor “model” with the covariate-membership relationship, DIF type, and DIF size were dependent on the third factor. Specifically, the interactions of the model and the relationship between the covariate and the latent variable were significant regardless of the DIF type with larger effect size for the non-DIF items, F = 18.342, p < .001, partial η^2 = .069, F = 27.242, p < .001, partial η^2 = .104. The interactions of the model and the DIF type were significant regardless of the number of DIF items included in the test with larger effect size when there were less DIF items, F = 19.038, p < .001, partial η^2 = .093, F = 15.432, p < .001, partial η^2 = .055. The interactions of the model and the DIF size were significant regardless of the group impact with larger effect for the impact group, F = 13.638, p < .001, partial η^2 = .045, F = 19.373, p < .001, partial η^2 = .070.

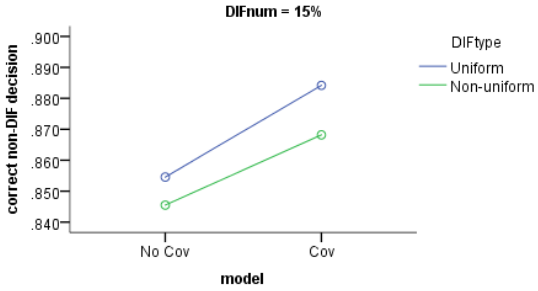
The interactions of the estimation method and the DIF type were significant regardless how strong the covariate was associated with the latent variable with larger effect size when the covariate-membership relationship was weak, F = 29.721, p < .001, partial η^2 = .118, F = 11.712, p < .001, partial η^2 = .034. The interactions of the estimation method and the DIF type were also dependent on the group impact with larger effect size for the impact group, F = 12.346, p < .001, partial η^2 = .042, F = 27.632, p < .001, partial η^2 = .114. The interaction of the estimation method and the group impact was significant only when a small number of DIF existed, F = 40.235, p < .001, partial η^2 = .136. Again, no interactions were found between the within-replication factors. The visual representations of these three-way interactions are shown in Figure 4.7.

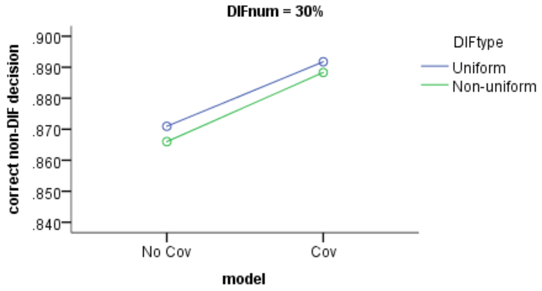
Table 4.8. Main and Interaction Effects of Factors on Correct Non-DIF Decision

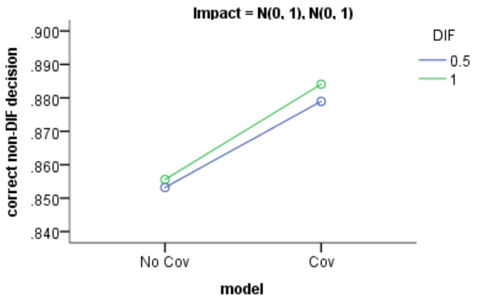
Source	**F value**	**p value**	**Partial η^2**
Within-replication factors:			
Model	79.133	<.001	.168
Model × OR	44.394	<.001	.103
Model × DIFtype	36.234	<.001	.073
Model × DIF	29.382	<.001	.066
Model × OR × DIFtype	41.532	<.001	.092
Model × DIFtype × DIFnum	37.439	<.001	.081
Model × DIF × Impact	27.394	<.001	.063
Estimation	83.493	<.001	.182
Estimation × DIFtype × OR	50.193	<.001	.113
Estimation × DIFtype × Impact	40.563	<.001	.090
Estimation × Impact × DIFnum	43.295	<.001	.102
Between-replication factors:			
OR	62.152	<.001	.122
DIFtype	31.752	<.001	.068
Impact	28.233	<.001	.066
Impact × DIFtype	53.832	<.001	.118
DIF × DIFtype × Impact	35.596	<.001	.073
DIFnum × DIFtype × DIF	42.483	<.001	.094

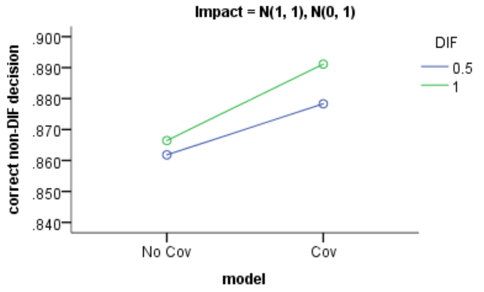

DIFtype = Uniform
correct non-DIF decision
.900
.890
.880
.870
.860
.850
.840
No Cov
Cov
model
OR
2
8

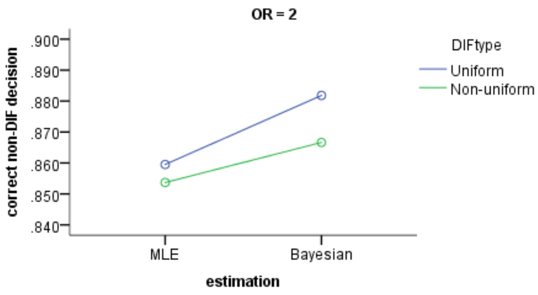

DIFtype = Non-uniform
correct non-DIF decision
.900
.890
.880
.870
.860
.850
.840
No Cov
Cov
model
OR
2
8

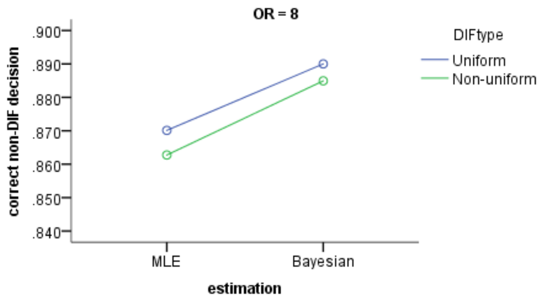

DIFnum = 15%
correct non-DIF decision
.900
.890
.880
.870
.860
.850
.840
No Cov
Cov
model
DIFtype
Uniform
Non-uniform

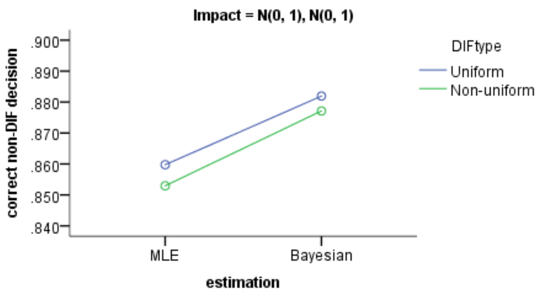

DIFnum = 30%
correct non-DIF decision
.900
.890
.880
.870
.860
.850
.840
No Cov
Cov
model
DIFtype
Uniform
Non-uniform

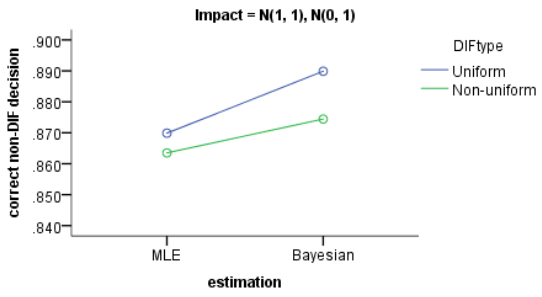

Impact = N(0, 1), N(0, 1)
correct non-DIF decision
.900
.890
.880
.870
.860
.850
.840
No Cov
Cov
model
DIF
0.5
1


Impact = N(1, 1), N(0, 1)
correct non-DIF decision
.900
.890
.880
.870
.860
.850
.840
No Cov
Cov
model
DIF
0.5
1


OR = 2
correct non-DIF decision
.900
.890
.880
.870
.860
.850
.840
MLE
Bayesian
estimation
DIFtype
Uniform
Non-uniform


OR = 8
correct non-DIF decision
.900
.890
.880
.870
.860
.850
.840
MLE
Bayesian
estimation
DIFtype
Uniform
Non-uniform


Impact = N(0, 1), N(0, 1)
correct non-DIF decision
.900
.890
.880
.870
.860
.850
.840
MLE
Bayesian
estimation
DIFtype
Uniform
Non-uniform


Impact = N(1, 1), N(0, 1)
correct non-DIF decision
.900
.890
.880
.870
.860
.850
.840
MLE
Bayesian
estimation
DIFtype
Uniform
Non-uniform

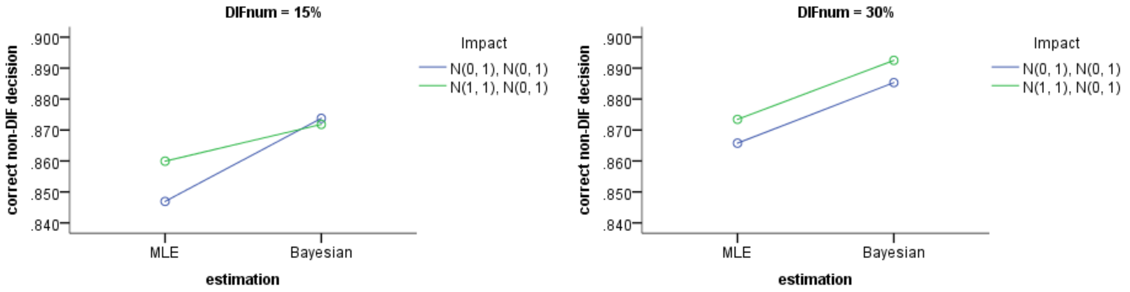


Figure 4.7. 3-Way Interactions of within-Replication Factors: Correct Non-DIF Decision

In line with the results for power, the interactions of the DIF number and the DIF type were dependent on the DIF magnitude with larger effect size when the DIF magnitude was small, F = 71.422, p < .001, η^2 = .110, F = 40.287, p < .001, η^2 = .069. In addition, the interactions of the DIF type and the DIF size were significant regardless of the group impact with larger effect size for the impact group, F = 29.224, p < .001, partial η^2 = .034, F = 56.473, p < .001, partial η^2 = .086. These interactions are presented in Figure 4.8.

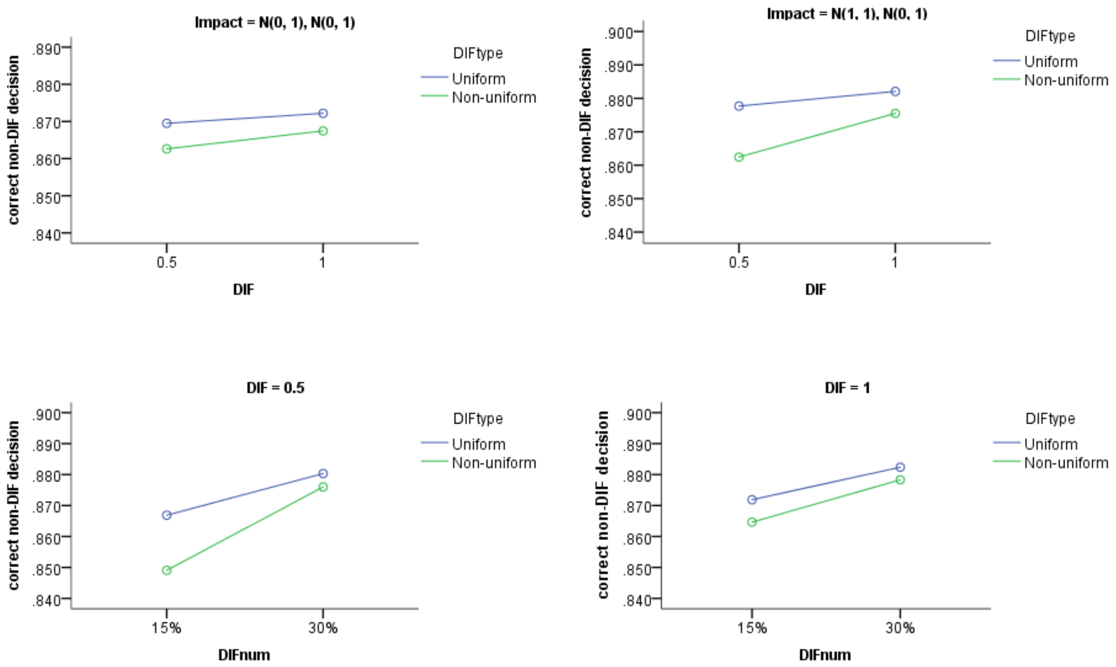


Figure 4.8. 3-Way Interactions of between-Replication Factors: Correct Non-DIF Decision

4.5.3 Correct Decision

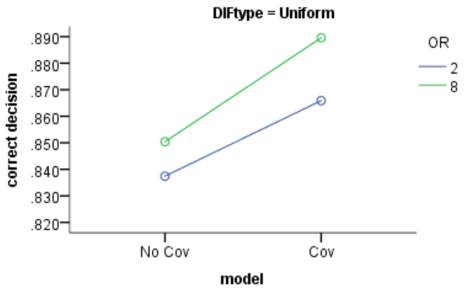
The measures of power and correct non-DIF decision were combined into a single index known as the "correct decision", which was used as a measure of the overall DIF detection efficiency. Table 4.9 summarizes the mixed ANOVA results of the manipulated factors on the correct decision.

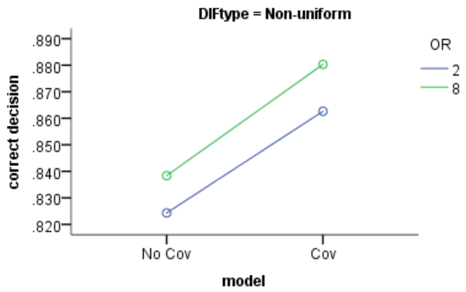
In line with the results for power and correct non-DIF decision, the interactions of the model and the association between the covariate and the latent group membership were significant for both the uniform and the non-uniform DIF items with larger effect size for the uniform DIF items, F = 20.125, p < .001, partial η^2 = .094, F = 16.225, p < .001, partial η^2 = .038. The interaction of the model and the DIF type was found to be significant only when a small number of DIF items existed, F = 19.783, p < .001, partial η^2 = .082. Likewise, the interaction of the model and the DIF size was significant only for the group impact condition, F = 28.325, p < .001, partial η^2 = .114.

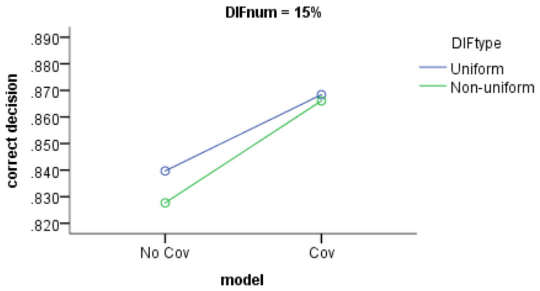
The interaction of the estimation method and the DIF type was found to be significant only when the covariate was weakly related to the latent group membership, F = 22.584, p < .001, partial η^2 = .090. The interaction of the estimation method and the DIF type was significant only when the group difference existed, F = 29.049, p < .001, partial η^2 = .115. Lastly, the interactions of the estimation method and the group impact were significant regardless of the number of DIF items, F = 17.245, p < .001, partial η^2 = .067, F = 12.831, p < .001, partial η^2 = .044. The visual representations of these three-way interactions involving the within-replication factors are shown in Figure 4.9.

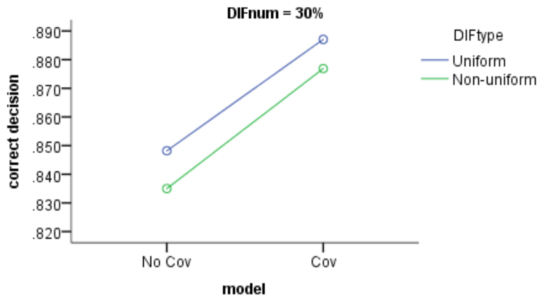
Table 4.9. Main and Interaction Effects of Factors on Correct Decision

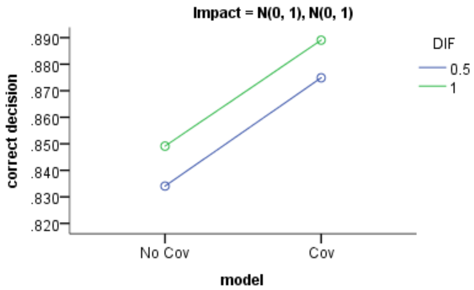
Source	**F value**	**p value**	**Partial η^2**
Within-replication factors:			
Model	42.342	<.001	.106
Model × OR	59.231	<.001	.145
Model × DIFtype	53.546	<.001	.128
Model × OR × DIFtype	39.421	<.001	.084
Model × DIFtype × DIFnum	35.765	<.001	.071
Model × DIF × Impact	45.582	<.001	.109
Estimation	60.325	<.001	.156
Estimation × DIFtype	44.203	<.001	.105
Estimation × Impact	38.127	<.001	.080
Estimation × DIFtype × OR	37.432	<.001	.074
Estimation × DIFtype × Impact	40.231	<.001	.086
Estimation × Impact × DIFnum	28.563	<.001	.065
Between-replication factors:			
OR	45.394	<.001	.104
DIFnum	29.573	<.001	.068
Impact	40.124	<.001	.088
DIFnum × DIF	46.432	<.001	.092
DIF × DIFtype	24.322	<.001	.062
DIF × DIFtype × Impact	38.439	<.001	.084
DIFnum × DIFtype × DIF	34.637	<.001	.077

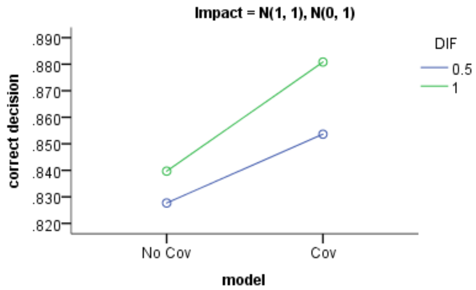

DIFtype = Uniform

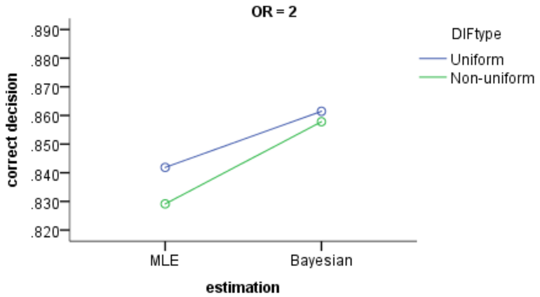

DIFtype = Non-uniform

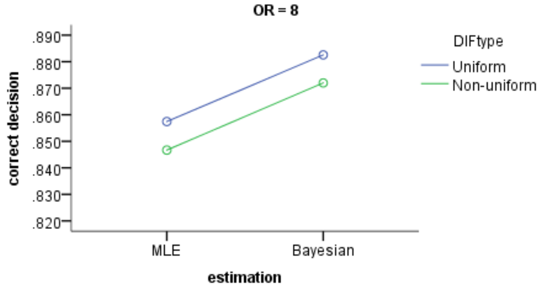

DIFnum = 15%


DIFnum = 30%


Impact = N(0, 1), N(0, 1)


Impact = N(1, 1), N(0, 1)


OR = 2


OR = 8

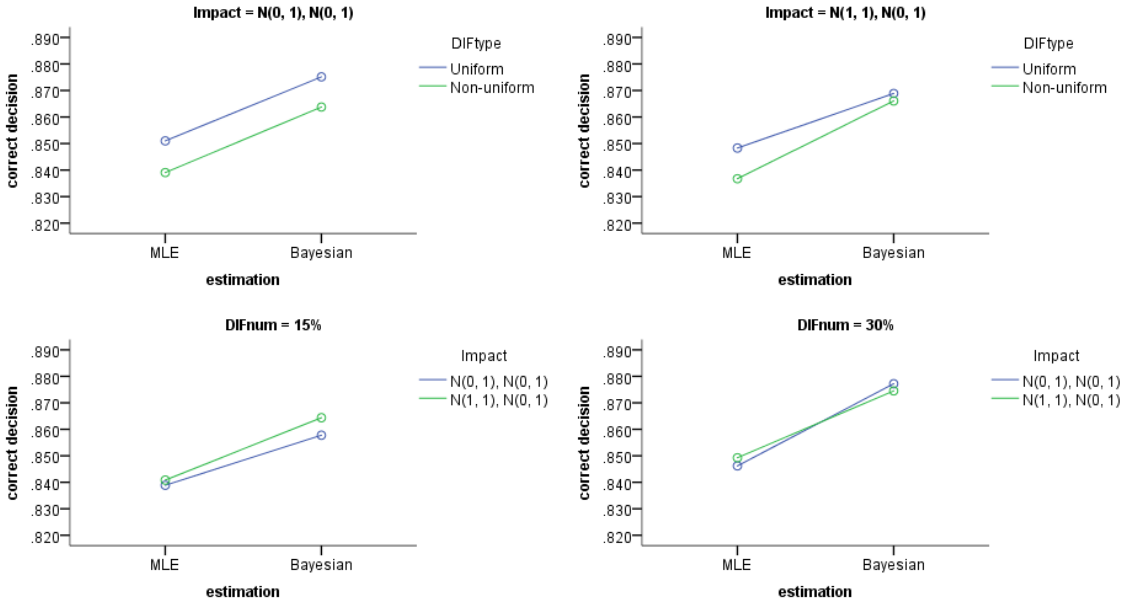
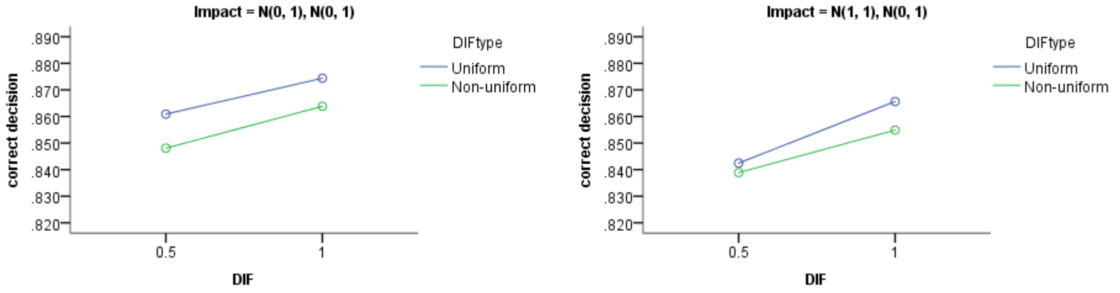


Figure 4.9. 3-Way Interactions of within-Replication Factors: Correct Decision

In line with the three-way interactions of between-replication factors for correct non-DIF decision, the interactions of the DIF number and the DIF type were dependent on the DIF magnitude with larger effect size when the DIF magnitude was small, F = 62.392, p < .001, η^2 = .098, F = 35.932, p < .001, η^2 = .044. The interactions of the DIF type and the DIF size were significant regardless of the group impact with larger effect size when the group difference existed, F = 32.932, p < .001, partial η^2 = .036, F = 59.296, p < .001, partial η^2 = .093. These interactions are presented in Figure 4.10.



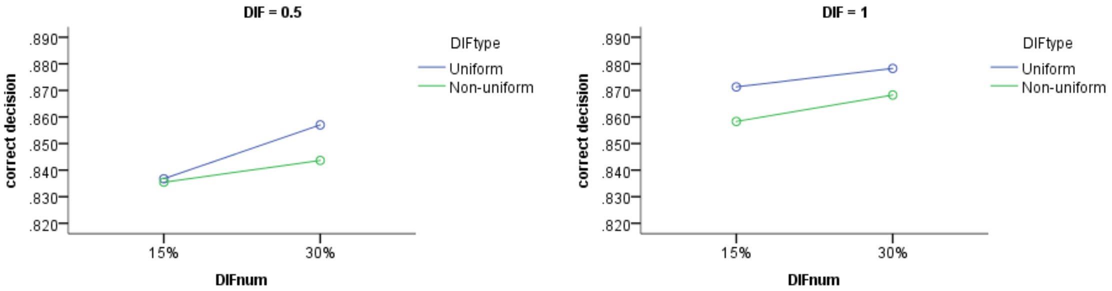


Figure 10. 3-Way Interactions of between-Replication Factors: Correct Decision

5.0 DISCUSSION

The focus of the present study is on the performance of mixture IRT models in latent DIF detection given the manipulation of relevant factors. The secondary purpose is to evaluate the recovery of the latent class structure and model parameters. This chapter summarizes the findings of the present study and discusses these findings by comparing among different outcome measures. Lastly, the limitations of the present study and the future directions are discussed.

5.1 DISCUSSION OF RESULTS

In this simulation study, the DIF size, the DIF magnitude, the proportion of DIF items, the group impact, the relationship between the covariate and the latent group membership, the inclusion of the covariate, and the estimation method were manipulated. Three research questions were proposed in the present study: 1) How do the simulation factors affect the performance of the mixture 2PL model in latent DIF detection? 2) How is the mixture modeling of latent DIF affected by the estimation method employed (MLE v.s. Bayesian estimation)? 3) How well are the latent DIF items detected under disadvantaged conditions.

The joint effects of the manipulated factors were focused on in the analysis. Overall, it was found that: 1) The relationship between the covariate and the latent group membership, the DIF type, and the proportion of DIF items played an important role in increasing the power of DIF detection and reducing the errors of model estimation. In addition, the group differences contributed to the DIF detection probably because it provided the information about group separation. 2) The Bayesian estimation in the detection of DIF with the 2PL model performed generally better than the MLE when the DIF items were uniform. Like the covariate effect, the group differences and the relationship between the covariate and the latent group membership affected the beneficial effect of the Bayesian estimation on DIF detection. 3) The use of the Bayesian estimation facilitated the detection of DIF in the situation where a fewer number of DIF items existed or the relationship between the covariate and the latent group membership was relatively weak. The Bayesian estimation also helped reduce the errors in the estimation of the model parameters with the mixture 2PL model. These results were discussed relative to each of the outcome measures below.

5.1.1 Recovery of Latent Structure

The analysis of the correct classification was performed as a standard procedure in the mixture IRT simulation literature. In the present study it was found that the highest classification accuracy was achieved when the DIF magnitude was large, sufficient DIF items were included, and the two groups differentiated in their abilities. Furthermore, consistent with the literature, it was found that the mixture 2PL model with the covariate and Bayesian estimation recovered the latent structure significantly better than the model without a covariate or using MLE. Interestingly, the latent group classification was significantly better when DIF items were uniform compared to the non-uniform DIF items. Consistent with previous studies, the average correct classification was moderate with the values ranging from .650 to .750. A previous simulation study found that the average correct classification was as high as above .900 (Li, 2014). This was probably due to two reasons. First, a higher DIF magnitude (i.e., $\Delta b = 1.5$) and higher covariate-membership association (i.e., OR = 10) were used. Second, a mixture Rasch model was used in that study. It is known that the classification accuracy is reduced as more complex models are used (Finch & French, 2012).

Previous literature (Smit et al., 1999, 2000) suggested that the latent class assignment substantially benefited from the incorporation of dichotomous covariates that were moderately or strongly associated with the latent class variable. This result was also found in the present study. Moreover, this pattern was demonstrated to be held regardless whether the DIF items were uniform or non-uniform. It was found that the uniform DIF items benefitted more from adding the covariate into the model regardless of how large the DIF size was or whether there was separation between the two latent groups.

In addition, group impact was found to be an important factor that determined the patterns of the covariate effect. For example, when there were fewer DIF items or smaller DIF magnitude, the covariate effect was found only when the group ability differences existed. This was probably because that the separation of the groups compensated for the information that was needed for latent class assignment.

Previous research suggested that the estimation methods play a role in mixture IRT modeling, and generally the MCMC algorithm produces more reasonable class classification accuracy. In line with the previous findings, the results of the present study showed the benefit of Bayesian estimation in class assignment when DIF size was small for both uniform and non-uniform DIF items. The benefit of Bayesian estimation was also displayed when there was no ability difference between the groups, however, it was only observed for uniform DIF items. More interestingly, there was an interaction between the estimation method and the covariate in that the covariate effect depended on the estimation method used. When less group information was available, the benefit of incorporating the covariate was larger when Bayesian estimation was used.

In contrast, when group differences appeared, the benefit of incorporating the covarite was smaller for Bayesian estimation. Even though the Bayesian estimation is notoriously known for a very lengthy period of time to complete a single analysis, it may facilitate the covariate effect in latent classification particularly under disadvantaged conditions when less group information is available in class assignment.

On the other hand, it was noted that the classification accuracy when using the mixture 2PL model depended largely on the relationship between the covariate and the latent variable. However, it was less dependent on the DIF size. Furthermore, the effect of the estimation method on the recovery of latent structure depended on the DIF size. However, it was less dependent on the number of DIF items, DIF size, or the relationship between the covariate and the latent variable.

5.1.2 Recovery of Model Parameters

RMSE has been used as a standard metric to evaluate model errors. It takes the square root of the average squared errors and thus penalizes variance as it gives the errors with larger absolute values more weight than the errors with smaller absolute values. It was used as a measure of the absolute errors in the recovery of the model parameters.

In line with the results for classification accuracy, the benefit of adding the covariate into the model was larger for uniform DIF items. This indicated that the correct latent class assignment improved with a reduction in parameter estimation errors, and this pattern was more pronounced for uniform DIF items. Furthermore, it suggests that covariate information may function differentially in model estimation and the benefits of the complex mixture IRT model may depend on the DIF type.

In general, the recovery of the model parameters benefited from the Bayesian estimation regardless of the nature of DIF. The reduction of errors when using the Bayesian estimation was more pronounced when the ability differences existed between the two latent groups. Regarding the benefits of using Bayesian estimation given disadvantaged conditions, it was found: 1) The reduction of errors when using the Bayesian estimation was larger when the relationship between the covariate and the latent variable or the DIF magnitude was small. 2) The reduction of error when using the Bayesian estimation was larger when the DIF size was small. 3) The reduction of errors when using the Bayesian estimation was larger for uniform DIF items compared to non-uniform DIF items particularly when the relationship between the covariate and the latent variable was small. These findings suggest that the use of MLE should probably be avoided when the association between the covariate and the latent group is unclear or the item DIF magnitude is small. This is because using MLE may bias the estimation of the mixture 2PL model and result in unsatisfactory latent class assignment.

5.1.3 Latent DIF Detection

Early studies showed that the latent class membership did not necessarily overlap with the manifest class membership, which brought up the necessity of studying latent DIF. In the analysis of latent DIF in the present study, an item was identified to be an DIF item only when it was identified to be significant by the logistic regression procedure and had non-negligible effect size. Three indices were then used to measure the appropriateness of the latent DIF detection. Power and correct non-DIF decision were equivalent to Type I and Type II error. When combined, power and correct non-DIF decision formed an overall measure of the DIF detection, termed correct decision in the present study.

In general, for most of the simulation conditions the power ranged from .700 to .850, and the correct non-DIF decision and the correct decision ranged from .850 to .900 and from .800 to .900, respectively. The power and correct decisions were moderately high to high. Increasing the number of replications may improve the detection of DIF using the mixture 2PL models. Using a set of unbiased anchor items may also increase the power of DIF detection (Lopez Rivas, Stark, & Chernyshenko, 2009; Wang, 2004).

In terms of the ANOVA results, the power of DIF detection was shown to rely on the DIF type, the proportion of DIF items, and the relationship between the covariate and the latent variable. Likewise, the correct decisions were also shown to rely on these factors. The benefit of adding the covariate into the model and using the Bayesian estimation had less dependence on the DIF magnitude or group impact.

Consistent interaction effects were found among the three indices of DIF detection regarding the covariate effect: 1) The covariate effect on the improving of DIF detection was better when the covariate was strongly correlated with the latent variable. 2) The covariate effect on improving the power of DIF detection was more significant for uniform DIF items, especially when the relationship between the covariate and the latent variable was weak. 3) The covariate effect on improving DIF detection was larger when the DIF magnitude was large, especially when the group differences existed.

The benefit of Bayesian estimation was also observed consistently among the three indices. Higher power and more correct decisions were achieved when the uniform DIF items were detected with the Bayesian estimation. This was more pronounced when the covariate was weakly related to the latent group membership and when the group differences existed. In addition, the facilitation of DIF detection for the impact group by using the Bayesian estimation depended on the number of DIF items. When a fewer number of DIF items were include, the detection of DIF was better for the impact group.

The above findings were generally consistent with the literature about the DIF detection with respect to the manifest group in the context of large-scale assessments. Svetina and Rutkowski (2014) found that the number of groups seemed to have no effect on the power or Type I error rates, whereas the magnitude of DIF, proportion of DIF items, and the nature/type of DIF affected the performance of DIF detection. It was found in the present study that the power was generally higher when the DIF magnitude was large or a large percentage of DIF items were included, which was expected. It was also found that the uniform DIF items were better distinguished compared to the non-uniform DIF items.

Lastly, the detection of DIF was better for the uniform DIF items when a large number of DIF items were included, and this was more pronounced when the DIF magnitude was small. Furthermore, the detection of DIF was better for the uniform DIF items when the DIF magnitude was large, and this was more pronounced for the impact group. These results suggest that the detection of DIF items relies on the separation of the latent groups and how much information about DIF is available for making the judgement. More interestingly, the type of DIF plays a role in DIF detection with mixture 2PL models.

5.2 LIMITATIONS AND FUTURE RESEARCH

The limitations of the present study included: 1) A relatively small number of replications were conducted due to the consideration of time required for model estimation under the Bayesian framework. 2) Anchor items were not included in the study design. 3) Some factors of interest were kept constant to reduce the complexity of the study design such as the proportion of covariate. It was fixed to 50/50, which was not a complete representation of the primary distributions of manifest groups in the population. 4) The detection of DIF was conducted after the latent group memberships were determined. The disadvantage was that the accuracy of DIF detection depended on the accuracy of latent group assignment. Alternatively, the analysis of DIF could be performed simultaneously with the model estimation in Mplus. Given that the MLE and the Bayesian estimation were compared, the post-hoc procedure was adopted to avoid the analysis of DIF using different estimation methods.

The study of DIF using mixture IRT models has been receiving more attention in recent years. Given some unavoidable issues in Bayesian estimation such as the label switching issue, the analysis of DIF with mixture IRT models was criticized and has not reached consistent conclusions. The present study shed the light on the future research regarding this topic. Given the dependence of DIF detection on DIF type discovered in the present study, researchers may consider examining the mechanisms that cause the differences in the detection of DIF between uniform and non-uniform DIF items. Future research could also consider incorporating more covariates into the model and vary the proportion of the manifest group membership to better understand the covariate effect on DIF detection.

To sum up, despite the limitations, the findings of the present study enriched the literature by expanding the understanding of the role of complex mixture IRT model in the detection of DIF. It provided the evidence regarding the benefits of the covariate and the Bayesian estimation in the identification of latent DIF. With an ever-increasing use of complicated models in psychometric practice, it is recommended to develop a good understanding of the nature of the test items and test takers before determining a relatively optimal analyzing model and estimation algorithm.

APPENDIX A

TABLES

Table A.1. Parameter Validation for Reference Latent Group

Item	Latent Group 1					
	a_{true}	$a_{estimated}$	$(a_{estimated} - a_{true})^2$	b_{true}	$b_{estimated}$	$(b_{estimated} - b_{true})^2$
1	0.840	0.825	0.000	0.915	0.878	0.001
2	1.163	1.129	0.001	-1.774	-1.771	0.000
3	1.429	1.504	0.006	-0.515	-0.571	0.003
4	0.520	0.426	0.009	1.401	1.740	0.115
5	0.713	0.764	0.003	1.083	0.982	0.010
6	1.140	1.143	0.000	0.864	0.916	0.003
7	0.825	0.807	0.000	1.546	1.445	0.010
8	0.788	0.806	0.000	-0.279	-0.376	0.009
9	1.167	1.205	0.001	1.556	1.620	0.004
10	1.336	1.450	0.013	-0.367	-0.328	0.002
11	0.446	0.468	0.000	0.899	0.791	0.012
12	0.670	0.615	0.003	-0.008	-0.031	0.001
13	0.994	1.009	0.000	-0.761	-0.784	0.001
14	0.688	0.707	0.000	1.123	1.106	0.000
15	1.357	1.436	0.006	-1.326	-1.327	0.000
16	0.662	0.559	0.011	0.269	0.463	0.038
17	0.459	0.425	0.001	0.565	0.287	0.077
18	1.787	1.849	0.004	0.251	0.188	0.004
19	2.031	2.076	0.002	-0.349	-0.432	0.007
20	1.737	1.751	0.000	-0.217	-0.197	0.000
21	1.579	1.532	0.002	-0.586	-0.589	0.000
22	0.541	0.612	0.005	-0.874	-0.693	0.033
23	1.974	1.881	0.009	-1.750	-1.864	0.013
24	0.659	0.631	0.001	-0.522	-0.549	0.001
25	0.794	0.744	0.003	0.142	-0.022	0.027
26	1.180	1.249	0.005	0.488	0.408	0.006
27	1.638	1.729	0.008	-1.044	-1.009	0.001
28	1.077	1.107	0.001	-0.832	-0.859	0.001
29	1.059	0.952	0.012	-1.570	-1.804	0.055

30	1.365	1.240	0.016	-0.949	-0.964	0.000
31	1.045	1.059	0.000	-0.510	-0.528	0.000
32	2.784	2.397	0.150	0.475	0.419	0.003
33	0.645	0.592	0.003	0.816	0.854	0.001
34	1.291	1.264	0.001	1.202	1.236	0.001
35	0.928	0.924	0.000	-0.674	-0.681	0.000
36	0.968	1.020	0.003	-1.223	-1.199	0.001
37	1.050	1.079	0.001	-0.299	-0.345	0.002
38	0.839	0.784	0.003	0.207	0.211	0.000
39	1.163	1.210	0.002	0.580	0.638	0.003
40	1.242	1.172	0.005	1.025	1.042	0.000

Table A.2. Parameter Validation for Focal Latent Group

Item	Latent Group 1					
	a_{true}	$a_{estimated}$	$(a_{estimated} - a_{true})^2$	b_{true}	$b_{estimated}$	$(b_{estimated} - b_{true})^2$
1	0.840	0.949	0.012	0.915	0.972	0.003
2	1.163	1.117	0.002	-1.774	-1.680	0.009
3	1.429	1.558	0.017	-0.515	-0.437	0.006
4	0.520	0.447	0.005	1.401	1.569	0.028
5	0.713	0.752	0.002	1.083	1.137	0.003
6	1.140	1.076	0.004	0.864	0.908	0.002
7	0.825	0.704	0.015	1.546	1.923	0.142
8	0.788	0.847	0.003	-0.279	-0.257	0.000
9	1.167	1.186	0.000	1.556	1.461	0.009
10	1.336	1.416	0.006	-0.367	-0.311	0.003
11	0.446	0.382	0.004	0.899	1.385	0.236
12	0.670	0.696	0.001	-0.008	-0.245	0.056
13	0.994	1.145	0.023	-0.761	-0.621	0.020
14	0.688	0.596	0.008	1.123	1.243	0.014
15	1.357	1.573	0.047	-1.326	-1.082	0.060
16	0.662	0.751	0.008	0.269	0.220	0.002
17	0.459	0.392	0.004	0.565	0.566	0.000
18	1.787	1.761	0.001	0.251	0.336	0.007
19	2.031	2.193	0.026	-0.349	-0.308	0.002
20	1.737	1.695	0.002	-0.217	-0.140	0.006
21	1.579	1.649	0.005	-0.586	-0.451	0.018
22	0.541	0.565	0.001	-0.874	-0.837	0.001
23	1.974	2.183	0.044	-1.750	-1.568	0.033
24	0.659	0.701	0.002	-0.522	-0.340	0.033
25	0.794	0.948	0.024	0.142	0.125	0.000
26	1.180	1.074	0.011	0.488	0.499	0.000
27	1.638	1.432	0.042	-1.044	-1.044	0.000
28	1.077	0.926	0.023	-0.832	-0.971	0.019
29	1.059	0.989	0.005	-1.570	-1.512	0.003
30	1.365	1.630	0.070	-0.949	-0.858	0.008
31	1.045	1.069	0.001	-0.510	-0.312	0.039
32	2.784	2.835	0.003	0.475	0.477	0.000
33	0.645	0.544	0.010	0.816	1.073	0.066
34	1.291	1.340	0.002	1.202	1.196	0.000
35	0.928	0.972	0.002	-0.174	-0.224	0.002
36	0.968	1.101	0.018	-0.723	-0.735	0.000
37	1.050	1.100	0.002	0.201	0.258	0.003
38	0.839	0.897	0.003	0.707	0.649	0.003
39	1.163	1.120	0.002	1.080	1.125	0.002

40	1.242	1.237	0.000	1.525	1.413	0.012

Table A.3. Example of the Generated Model Parameters for Uniform DIF

Item	Δb = 0.5, DIFnum = 6				Δb = 0.5, DIFnum = 12			
	Reference Latent Group		Focal Latent Group		Reference Latent Group		Focal Latent Group	
	a	b	a	b	a	b	a	b
1	0.840	0.915	0.840	0.915	0.840	0.915	0.840	0.915
2	1.163	-1.774	1.163	-1.774	1.163	-1.774	1.163	-1.774
3	1.429	-0.515	1.429	-0.515	1.429	-0.515	1.429	-0.515
4	0.520	1.401	0.520	1.401	0.520	1.401	0.520	1.401
5	0.713	1.083	0.713	1.083	0.713	1.083	0.713	1.083
6	1.140	0.864	1.140	0.864	1.140	0.864	1.140	0.864
7	0.825	1.546	0.825	1.546	0.825	1.546	0.825	1.546
8	0.788	-0.279	0.788	-0.279	0.788	-0.279	0.788	-0.279
9	1.167	1.556	1.167	1.556	1.167	1.556	1.167	1.556
10	1.336	-0.367	1.336	-0.367	1.336	-0.367	1.336	-0.367
11	0.446	0.899	0.446	0.899	0.446	0.899	0.446	0.899
12	0.670	-0.008	0.670	-0.008	0.670	-0.008	0.670	-0.008
13	0.994	-0.761	0.994	-0.761	0.994	-0.761	0.994	-0.761
14	0.688	1.123	0.688	1.123	0.688	1.123	0.688	1.123
15	1.357	-1.326	1.357	-1.326	1.357	-1.326	1.357	-1.326
16	0.662	0.269	0.662	0.269	0.662	0.269	0.662	0.269
17	0.459	0.565	0.459	0.565	0.459	0.565	0.459	0.565
18	1.787	0.251	1.787	0.251	1.787	0.251	1.787	0.251
19	2.031	-0.349	2.031	-0.349	2.031	-0.349	2.031	-0.349
20	1.737	-0.217	1.737	-0.217	1.737	-0.217	1.737	-0.217
21	1.579	-0.586	1.579	-0.586	1.579	-0.586	1.579	-0.586
22	0.541	-0.874	0.541	-0.874	0.541	-0.874	0.541	-0.874
23	1.974	-1.750	1.974	-1.750	1.974	-1.750	1.974	-1.750
24	0.659	-0.522	0.659	-0.522	0.659	-0.522	0.659	-0.522
25	0.794	0.142	0.794	0.142	0.794	0.142	0.794	0.142
26	1.180	0.488	1.180	0.488	1.180	0.488	1.180	0.488
27	1.638	-1.044	1.638	-1.044	1.638	-1.044	1.638	-1.044
28	1.077	-0.832	1.077	-0.832	1.077	-0.832	1.077	-0.832
29	1.059	-1.570	1.059	-1.570	1.059	-1.570	1.059	-1.070
30	1.365	-0.949	1.365	-0.949	1.365	-0.949	1.365	-0.449
31	1.045	-0.510	1.045	-0.510	1.045	-0.510	1.045	-0.010
32	2.784	0.475	2.784	0.475	2.784	0.475	2.784	0.975
33	0.645	0.816	0.645	0.816	0.645	0.816	0.645	1.316
34	1.291	1.202	1.291	1.202	1.291	1.202	1.291	1.702
35	0.928	-0.674	0.928	-0.174	0.928	-0.674	0.928	-0.174
36	0.968	-1.223	0.968	-0.723	0.968	-1.223	0.968	-0.723
37	1.050	-0.299	1.050	0.201	1.050	-0.299	1.050	0.201
38	0.839	0.207	0.839	0.707	0.839	0.207	0.839	0.707

39	1.163	0.580	1.163	1.080	1.163	0.580	1.163	1.080
40	1.242	1.025	1.242	1.525	1.242	1.025	1.242	1.525

Table A.4. Example of the Generated Model Parameters for Non-uniform DIF

Item	Δa = 1, DIFnum = 6				Δa = 1, DIFnum = 12			
	Reference Latent Group		Focal Latent Group		Reference Latent Group		Focal Latent Group	
	a	b	a	b	a	b	a	b
1	0.840	0.915	0.840	0.915	0.840	0.915	0.840	0.915
2	1.163	-1.774	1.163	-1.774	1.163	-1.774	1.163	-1.774
3	0.520	1.401	0.520	1.401	0.520	1.401	0.520	1.401
4	0.713	1.083	0.713	1.083	0.713	1.083	0.713	1.083
5	1.140	0.864	1.140	0.864	1.140	0.864	1.140	0.864
6	0.825	1.546	0.825	1.546	0.825	1.546	0.825	1.546
7	0.788	-0.279	0.788	-0.279	0.788	-0.279	0.788	-0.279
8	1.167	1.556	1.167	1.556	1.167	1.556	1.167	1.556
9	0.446	0.899	0.446	0.899	0.446	0.899	0.446	0.899
10	0.670	-0.008	0.670	-0.008	0.670	-0.008	0.670	-0.008
11	0.994	-0.761	0.994	-0.761	0.994	-0.761	0.994	-0.761
12	0.688	1.123	0.688	1.123	0.688	1.123	0.688	1.123
13	0.662	0.269	0.662	0.269	0.662	0.269	0.662	0.269
14	0.459	0.565	0.459	0.565	0.459	0.565	0.459	0.565
15	0.541	-0.874	0.541	-0.874	0.541	-0.874	0.541	-0.874
16	0.659	-0.522	0.659	-0.522	0.659	-0.522	0.659	-0.522
17	0.794	0.142	0.794	0.142	0.794	0.142	0.794	0.142
18	1.180	0.488	1.180	0.488	1.180	0.488	1.180	0.488
19	1.077	-0.832	1.077	-0.832	1.077	-0.832	1.077	-0.832
20	1.059	-1.570	1.059	-1.570	1.059	-1.570	1.059	-1.570
21	1.045	-0.510	1.045	-0.510	1.045	-0.510	1.045	-0.510
22	0.645	0.816	0.645	0.816	0.645	0.816	0.645	0.816
23	0.928	-0.674	0.928	-0.674	0.928	-0.674	0.928	-0.674
24	0.968	-1.223	0.968	-1.223	0.968	-1.223	0.968	-1.223
25	1.050	-0.299	1.050	-0.299	1.050	-0.299	1.050	-0.299
26	0.839	0.207	0.839	0.207	0.839	0.207	0.839	0.207
27	1.163	0.580	1.163	0.580	1.163	0.580	1.163	0.580
28	1.242	1.025	1.242	1.025	1.242	1.025	1.242	1.025
29	1.291	1.202	1.291	1.202	1.291	1.202	0.291	1.202
30	1.336	-0.367	1.336	-0.367	1.336	-0.367	0.336	-0.367
31	1.357	-1.326	1.357	-1.326	1.357	-1.326	0.357	-1.326
32	1.365	-0.949	1.365	-0.949	1.365	-0.949	0.365	-0.949
33	1.429	-0.515	1.429	-0.515	1.429	-0.515	0.429	-0.515
34	2.784	0.475	2.784	0.475	2.784	0.475	1.784	0.475
35	1.579	-0.586	0.579	-0.586	1.579	-0.586	0.579	-0.586
36	1.638	-1.044	0.638	-1.044	1.638	-1.044	0.638	-1.044
37	1.737	-0.217	0.737	-0.217	1.737	-0.217	0.737	-0.217
38	1.787	0.251	0.787	0.251	1.787	0.251	0.787	0.251

39	1.974	-1.750	0.974	-1.750	1.974	-1.750	0.974	-1.750
40	2.031	-0.349	1.031	-0.349	2.031	-0.349	1.031	-0.349

Table A.5. Convergence Results for Bayesian Estimation

Results	Results Type	Percent	Within-chain label switching	Between-chain label switching	Collapsed	Recovered	Poor mixing
Two chains converged	Within-chain label switching	0.642%	0.642%				
	Collapsed	9.151%			9.151%		
	Recovered	69.543%				69.543%	
	Poor mixing	1.277%					1.277%
Two chains non-converged	Within-chain label switching	0.385%	0.385%				
	Between-chain label switching	10.623%		10.623%			
	Collapsed	5.169%			5.169%		
	Recovered	2.969%				2.969%	
	Poor mixing	0.241%					0.241%
Total		100%	1.027%	10.623%	14.320%	72.512%	1.518%

Table A.6. Average Latent Group Classification Accuracy of Each Simulation Condition

					NoCov				**Cov**			
					MLE		**Bayesian**		**MLE**		**Bayesian**	
OR	**DIFtype**	**DIFnum**	**DIF**	**Impact**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
2	Uniform	15%	0.5	No	0.569	0.137	0.607	0.119	0.573	0.105	0.619	0.130
				Yes	0.530	0.115	0.603	0.148	0.614	0.130	0.691	0.049
			1	No	0.736	0.026	0.731	0.051	0.738	0.046	0.746	0.021
				Yes	0.734	0.042	0.776	0.024	0.780	0.023	0.807	0.016
		30%	0.5	No	0.629	0.072	0.575	0.138	0.621	0.106	0.640	0.094
				Yes	0.583	0.125	0.666	0.115	0.640	0.113	0.712	0.040
			1	No	0.826	0.010	0.815	0.009	0.827	0.011	0.824	0.012
				Yes	0.863	0.019	0.871	0.013	0.874	0.013	0.876	0.014
	Non-uniform	15%	0.5	No	0.523	0.146	0.587	0.132	0.536	0.141	0.599	0.089
				Yes	0.518	0.152	0.616	0.126	0.528	0.142	0.677	0.031
			1	No	0.652	0.075	0.630	0.090	0.648	0.085	0.683	0.041
				Yes	0.715	0.010	0.709	0.007	0.727	0.014	0.710	0.009
		30%	0.5	No	0.599	0.110	0.668	0.051	0.564	0.128	0.667	0.051
				Yes	0.648	0.092	0.668	0.017	0.671	0.014	0.665	0.037
			1	No	0.808	0.014	0.799	0.017	0.812	0.011	0.807	0.019
				Yes	0.836	0.011	0.843	0.010	0.843	0.010	0.847	0.010
8	Uniform	15%	0.5	No	0.593	0.121	0.543	0.155	0.645	0.115	0.626	0.115
				Yes	0.480	0.120	0.668	0.087	0.763	0.027	0.770	0.020
			1	No	0.742	0.017	0.731	0.024	0.794	0.019	0.807	0.011
				Yes	0.701	0.070	0.775	0.028	0.850	0.009	0.855	0.008
		30%	0.5	No	0.586	0.136	0.598	0.130	0.715	0.066	0.705	0.050
				Yes	0.601	0.133	0.640	0.144	0.785	0.034	0.795	0.036
			1	No	0.817	0.019	0.825	0.013	0.861	0.009	0.862	0.007
				Yes	0.840	0.035	0.866	0.011	0.900	0.008	0.896	0.006
	Non-uniform	15%	0.5	No	0.531	0.124	0.587	0.109	0.523	0.137	0.629	0.082
				Yes	0.519	0.140	0.680	0.019	0.701	0.042	0.720	0.018
			1	No	0.620	0.100	0.645	0.104	0.702	0.086	0.730	0.015
				Yes	0.711	0.007	0.711	0.008	0.769	0.026	0.772	0.033
		30%	0.5	No	0.619	0.100	0.583	0.134	0.597	0.141	0.645	0.094
				Yes	0.615	0.129	0.665	0.026	0.635	0.128	0.665	0.029

1	No	0.805	0.010	0.805	0.016	0.840	0.007	0.836	0.011
	Yes	0.834	0.013	0.839	0.008	0.868	0.009	0.873	0.007

Table A.7. Average RMSE of Each Simulation Condition

OR	DIFtype	DIFnum	DIF	Impact	NoCov				Cov			
					MLE		Bayesian		MLE		Bayesian	
					M	SD	M	SD	M	SD	M	SD
2	Uniform	15%	0.5	No	0.256	0.003	0.248	0.003	0.258	0.003	0.249	0.005
				Yes	0.331	0.027	0.269	0.015	0.284	0.017	0.249	0.004
			1	No	0.271	0.013	0.249	0.004	0.263	0.006	0.248	0.004
				Yes	0.303	0.023	0.255	0.014	0.265	0.010	0.243	0.005
		30%	0.5	No	0.265	0.008	0.248	0.004	0.261	0.007	0.248	0.004
				Yes	0.332	0.031	0.268	0.016	0.296	0.026	0.253	0.008
			1	No	0.256	0.005	0.246	0.003	0.258	0.005	0.247	0.003
				Yes	0.249	0.003	0.240	0.005	0.248	0.003	0.240	0.006
	Non-uniform	15%	0.5	No	0.260	0.004	0.250	0.003	0.260	0.004	0.251	0.003
				Yes	0.313	0.011	0.275	0.014	0.314	0.020	0.264	0.011
			1	No	0.264	0.003	0.257	0.003	0.264	0.003	0.255	0.003
				Yes	0.257	0.003	0.249	0.002	0.257	0.003	0.249	0.003
		30%	0.5	No	0.263	0.004	0.253	0.003	0.265	0.005	0.254	0.003
				Yes	0.267	0.008	0.253	0.004	0.265	0.008	0.253	0.004
			1	No	0.267	0.001	0.261	0.003	0.267	0.001	0.261	0.003
				Yes	0.261	0.002	0.256	0.005	0.262	0.002	0.257	0.001
8	Uniform	15%	0.5	No	0.260	0.003	0.249	0.004	0.256	0.003	0.248	0.003
				Yes	0.341	0.029	0.265	0.011	0.254	0.004	0.245	0.002
			1	No	0.270	0.010	0.249	0.004	0.255	0.002	0.247	0.004
				Yes	0.314	0.031	0.254	0.009	0.250	0.002	0.240	0.004
		30%	0.5	No	0.260	0.007	0.249	0.003	0.256	0.005	0.249	0.002
				Yes	0.324	0.034	0.265	0.015	0.251	0.002	0.242	0.006
			1	No	0.260	0.008	0.248	0.004	0.257	0.003	0.249	0.004
				Yes	0.260	0.010	0.241	0.003	0.249	0.002	0.241	0.005
	Non-uniform	15%	0.5	No	0.263	0.006	0.254	0.004	0.265	0.006	0.253	0.005
				Yes	0.335	0.019	0.263	0.010	0.262	0.011	0.245	0.007
			1	No	0.265	0.002	0.256	0.004	0.263	0.003	0.254	0.003
				Yes	0.255	0.002	0.247	0.003	0.256	0.003	0.247	0.003
		30%	0.5	No	0.265	0.004	0.254	0.003	0.262	0.003	0.254	0.003
				Yes	0.267	0.006	0.252	0.005	0.271	0.009	0.253	0.005

1	No	0.266	0.002	0.259	0.003	0.264	0.002	0.259	0.004
	Yes	0.266	0.002	0.256	0.005	0.263	0.002	0.255	0.003

Table A.8. Average Power of Each Simulation Condition

					NoCov				Cov			
					MLE		Bayesian		MLE		Bayesian	
OR	DIFtype	DIFnum	DIF	Impact	M	SD	M	SD	M	SD	M	SD
2	Uniform	15%	0.5	No	0.632	0.062	0.666	0.122	0.720	0.117	0.734	0.057
				Yes	0.649	0.058	0.677	0.088	0.723	0.119	0.735	0.084
			1	No	0.635	0.071	0.702	0.088	0.759	0.102	0.762	0.059
				Yes	0.647	0.088	0.694	0.081	0.797	0.069	0.806	0.101
		30%	0.5	No	0.664	0.093	0.699	0.081	0.727	0.097	0.742	0.117
				Yes	0.678	0.098	0.740	0.115	0.756	0.097	0.804	0.083
			1	No	0.672	0.097	0.731	0.097	0.788	0.083	0.833	0.061
				Yes	0.695	0.081	0.782	0.106	0.829	0.045	0.898	0.055
	Non-uniform	15%	0.5	No	0.646	0.055	0.708	0.081	0.653	0.093	0.769	0.119
				Yes	0.653	0.077	0.699	0.116	0.653	0.123	0.778	0.105
			1	No	0.643	0.090	0.700	0.098	0.760	0.138	0.800	0.086
				Yes	0.665	0.098	0.735	0.150	0.803	0.041	0.859	0.048
		30%	0.5	No	0.677	0.043	0.745	0.118	0.825	0.091	0.870	0.116
				Yes	0.710	0.081	0.792	0.053	0.841	0.088	0.893	0.055
			1	No	0.688	0.055	0.772	0.072	0.856	0.098	0.910	0.104
				Yes	0.714	0.048	0.828	0.071	0.734	0.136	0.943	0.081
8	Uniform	15%	0.5	No	0.642	0.090	0.664	0.098	0.748	0.095	0.797	0.120
				Yes	0.643	0.111	0.671	0.093	0.757	0.079	0.811	0.062
			1	No	0.665	0.093	0.680	0.071	0.778	0.055	0.811	0.045
				Yes	0.678	0.119	0.698	0.081	0.806	0.117	0.864	0.069
		30%	0.5	No	0.689	0.047	0.713	0.097	0.826	0.054	0.900	0.044
				Yes	0.738	0.111	0.759	0.055	0.847	0.055	0.912	0.033
			1	No	0.742	0.037	0.810	0.092	0.857	0.048	0.920	0.027
				Yes	0.790	0.093	0.831	0.119	0.881	0.111	0.964	0.048
	Non-uniform	15%	0.5	No	0.645	0.105	0.687	0.081	0.790	0.105	0.818	0.079
				Yes	0.663	0.056	0.716	0.070	0.803	0.081	0.825	0.105
			1	No	0.671	0.039	0.721	0.055	0.792	0.055	0.889	0.076
				Yes	0.685	0.028	0.739	0.044	0.818	0.071	0.911	0.117
		30%	0.5	No	0.707	0.071	0.757	0.080	0.841	0.044	0.926	0.099
				Yes	0.768	0.121	0.866	0.075	0.907	0.055	0.935	0.070

	1	No	0.813	0.024	0.855	0.046	0.896	0.069	0.955	0.058
		Yes	0.834	0.046	0.921	0.037	0.935	0.056	0.972	0.034

Table A.9. Average Correct Non-DIF Decision of Each Simulation Condition

					NoCov				**Cov**			
					MLE		**Bayesian**		**MLE**		**Bayesian**	
OR	**DIFtype**	**DIFnum**	**DIF**	**Impact**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
2	Uniform	15%	0.5	No	0.828	0.028	0.849	0.037	0.854	0.033	0.890	0.034
				Yes	0.840	0.037	0.855	0.042	0.866	0.026	0.896	0.027
			1	No	0.830	0.032	0.848	0.020	0.857	0.023	0.894	0.024
				Yes	0.852	0.028	0.864	0.024	0.872	0.033	0.902	0.034
		30%	0.5	No	0.854	0.033	0.869	0.024	0.870	0.031	0.896	0.032
				Yes	0.864	0.023	0.880	0.024	0.882	0.025	0.904	0.026
			1	No	0.856	0.033	0.874	0.023	0.877	0.029	0.900	0.030
				Yes	0.868	0.035	0.882	0.020	0.882	0.033	0.906	0.034
	Non-uniform	15%	0.5	No	0.820	0.028	0.835	0.029	0.840	0.040	0.882	0.041
				Yes	0.832	0.038	0.850	0.024	0.868	0.035	0.729	0.036
			1	No	0.826	0.021	0.842	0.027	0.848	0.024	0.892	0.025
				Yes	0.842	0.024	0.856	0.024	0.864	0.036	0.899	0.037
		30%	0.5	No	0.850	0.040	0.868	0.031	0.866	0.037	0.890	0.038
				Yes	0.859	0.019	0.874	0.040	0.876	0.019	0.900	0.020
			1	No	0.854	0.026	0.870	0.027	0.872	0.025	0.900	0.026
				Yes	0.862	0.026	0.876	0.027	0.880	0.026	0.903	0.027
8	Uniform	15%	0.5	No	0.846	0.025	0.868	0.030	0.872	0.025	0.896	0.026
				Yes	0.855	0.038	0.873	0.031	0.876	0.021	0.906	0.022
			1	No	0.848	0.038	0.872	0.033	0.876	0.026	0.900	0.027
				Yes	0.865	0.040	0.880	0.024	0.884	0.025	0.906	0.026
		30%	0.5	No	0.868	0.028	0.874	0.028	0.878	0.024	0.900	0.025
				Yes	0.872	0.030	0.882	0.033	0.884	0.019	0.908	0.020
			1	No	0.862	0.025	0.876	0.033	0.880	0.023	0.905	0.028
				Yes	0.870	0.029	0.884	0.034	0.886	0.029	0.910	0.030
	Non-uniform	15%	0.5	No	0.834	0.027	0.858	0.024	0.866	0.021	0.890	0.022
				Yes	0.846	0.027	0.862	0.020	0.873	0.026	0.900	0.027
			1	No	0.838	0.019	0.866	0.020	0.868	0.027	0.898	0.025
				Yes	0.852	0.020	0.869	0.027	0.872	0.027	0.902	0.028
		30%	0.5	No	0.858	0.019	0.872	0.020	0.875	0.020	0.898	0.027
				Yes	0.866	0.021	0.879	0.020	0.880	0.019	0.905	0.020

	1	No	0.854	0.023	0.873	0.024	0.878	0.021	0.900	0.021
		Yes	0.860	0.017	0.881	0.022	0.884	0.101	0.906	0.020

Table A.10. Average Correct Decision of Each Simulation Condition

OR	DIFtype	DIFnum	DIF	Impact	NoCov				Cov			
					MLE		Bayesian		MLE		Bayesian	
					M	SD	M	SD	M	SD	M	SD
2	Uniform	15%	0.5	No	0.799	0.029	0.822	0.037	0.834	0.031	0.867	0.033
				Yes	0.811	0.035	0.828	0.034	0.845	0.033	0.872	0.032
			1	No	0.801	0.024	0.826	0.026	0.842	0.034	0.874	0.029
				Yes	0.821	0.034	0.839	0.034	0.861	0.033	0.888	0.030
		30%	0.5	No	0.826	0.040	0.844	0.027	0.849	0.029	0.873	0.030
				Yes	0.836	0.025	0.859	0.021	0.863	0.029	0.889	0.020
			1	No	0.828	0.031	0.853	0.032	0.864	0.027	0.890	0.022
				Yes	0.842	0.040	0.867	0.022	0.874	0.038	0.905	0.032
	Non-uniform	15%	0.5	No	0.794	0.023	0.816	0.033	0.812	0.039	0.865	0.027
				Yes	0.805	0.035	0.830	0.028	0.836	0.034	0.736	0.042
			1	No	0.799	0.025	0.823	0.024	0.835	0.034	0.878	0.029
				Yes	0.815	0.028	0.838	0.020	0.855	0.054	0.893	0.029
		30%	0.5	No	0.824	0.039	0.850	0.028	0.860	0.036	0.887	0.030
				Yes	0.837	0.025	0.862	0.033	0.871	0.028	0.899	0.040
			1	No	0.829	0.030	0.855	0.027	0.870	0.043	0.902	0.052
				Yes	0.840	0.029	0.869	0.043	0.858	0.036	0.909	0.053
8	Uniform	15%	0.5	No	0.815	0.028	0.837	0.033	0.853	0.025	0.881	0.032
				Yes	0.823	0.038	0.843	0.026	0.858	0.034	0.892	0.031
			1	No	0.821	0.027	0.843	0.044	0.861	0.033	0.887	0.027
				Yes	0.837	0.035	0.853	0.024	0.872	0.027	0.900	0.046
		30%	0.5	No	0.841	0.030	0.850	0.029	0.870	0.021	0.900	0.028
				Yes	0.852	0.025	0.864	0.032	0.878	0.024	0.909	0.039
			1	No	0.844	0.028	0.866	0.038	0.877	0.030	0.907	0.036
				Yes	0.858	0.033	0.876	0.038	0.885	0.031	0.918	0.037
	Non-uniform	15%	0.5	No	0.806	0.025	0.832	0.029	0.855	0.030	0.879	0.037
				Yes	0.819	0.026	0.838	0.022	0.863	0.037	0.889	0.042
			1	No	0.813	0.025	0.841	0.027	0.857	0.034	0.897	0.034
				Yes	0.827	0.027	0.846	0.026	0.864	0.035	0.903	0.049
		30%	0.5	No	0.835	0.028	0.852	0.027	0.870	0.027	0.902	0.034
				Yes	0.851	0.027	0.877	0.027	0.884	0.036	0.910	0.020

1	No	0.848	0.035	0.870	0.029	0.881	0.032	0.908	0.035
	Yes	0.856	0.033	0.887	0.039	0.892	0.043	0.916	0.037

APPENDIX B

RESPONSE DATA GENERATION IN SAS

```
/* F1: OR */
%let Lx=372;
%let Hx=492;

**set macro loop;
%macro datasim;
/*read in 2PL a and b parameters */

proc iml;

x={0.83962 1.16344 1.42861 0.51953 0.71311 1.14013 0.82456 0.78848 1.16716
1.33606 0.44606 0.67009 0.99443 0.68752 1.35674 0.66216 0.45865 1.78664
2.03085 1.73668 1.57925 0.54061 1.97391 0.65948 0.79420 1.18033 1.63773
1.07664 1.05928 1.36463 1.04468 2.78423 0.64483 1.29112 0.92831 0.96820
1.05021 0.83922 1.16335 1.24204 0.91476 -1.77352 -0.51495 1.40075 1.08319
0.86375 1.54594 -0.27873 1.55580 -0.36707 0.89879 -0.00758 -0.76079 1.12282 -
1.32595 0.26900 0.56473 0.25081 -0.34929 -0.21656 -0.58608 -0.87356 -1.75013
-0.52190 0.14161 0.48799 -1.04382 -0.83242 -1.56983 -0.94872 -0.51035 0.47500
0.81623 1.20175 -0.67405 -1.22326 -0.29884 0.20706 0.57961 1.02461};
create item_para1 from x[colname={"a1" "a2" "a3" "a4" "a5" "a6" "a7" "a8"
"a9" "a10" "a11" "a12" "a13" "a14" "a15" "a16" "a17" "a18" "a19" "a20" "a21"
"a22" "a23" "a24" "a25" "a26" "a27" "a28" "a29" "a30" "a31" "a32" "a33" "a34"
"a35" "a36" "a37" "a38" "a39" "a40" "b1" "b2" "b3" "b4" "b5" "b6" "b7" "b8"
"b9" "b10" "b11" "b12" "b13" "b14" "b15" "b16" "b17" "b18" "b19" "b20" "b21"
"b22" "b23" "b24" "b25" "b26" "b27" "b28" "b29" "b30" "b31" "b32" "b33" "b34"
"b35" "b36" "b37" "b38" "b39" "b40"}];
append from x;
close item_para1;

y={0.83962 1.16344 1.42861 0.51953 0.71311 1.14013 0.82456 0.78848 1.16716
1.33606 0.44606 0.67009 0.99443 0.68752 1.35674 0.66216 0.45865 1.78664
2.03085 1.73668 1.57925 0.54061 1.97391 0.65948 0.79420 1.18033 1.63773
1.07664 1.05928 1.36463 1.04468 2.78423 0.64483 1.29112 0.92831 0.96820
1.05021 0.83922 1.16335 1.24204 0.91476 -1.77352 -0.51495 1.40075 1.08319
0.86375 1.54594 -0.27873 1.55580 -0.36707 0.89879 -0.00758 -0.76079 1.12282 -
1.32595 0.26900 0.56473 0.25081 -0.34929 -0.21656 -0.58608 -0.87356 -1.75013
-0.52190 0.14161 0.48799 -1.04382 -0.83242 -1.56983 -0.94872 -0.51035 0.47500
0.81623 1.20175 -0.17405 -0.72326 0.20116 0.70706 1.07961 1.52461};
create item_para2 from y[colname={"a1" "a2" "a3" "a4" "a5" "a6" "a7" "a8"
"a9" "a10" "a11" "a12" "a13" "a14" "a15" "a16" "a17" "a18" "a19" "a20" "a21"
"a22" "a23" "a24" "a25" "a26" "a27" "a28" "a29" "a30" "a31" "a32" "a33" "a34"
"a35" "a36" "a37" "a38" "a39" "a40" "b1" "b2" "b3" "b4" "b5" "b6" "b7" "b8"
"b9" "b10" "b11" "b12" "b13" "b14" "b15" "b16" "b17" "b18" "b19" "b20" "b21"
"b22" "b23" "b24" "b25" "b26" "b27" "b28" "b29" "b30" "b31" "b32" "b33" "b34"
"b35" "b36" "b37" "b38" "b39" "b40"}];
append from y;
close item_para2; quit;

%do ii=1 %to 2;
%put "Starting Loop &ii";
%do f1=1 %to 2;
      %if &f1=1 %then %let x=&Lx;
      %if &f1=2 %then %let x=&Hx;

%let OR=%eval(&f1);
%let seed=%eval(10000*&f1+1000*&ii);

/* generate response data */

data respU1LC1;
set item_para1;

array resp(*) resp1-resp40;
array a(*) a1-a40;
array b(*) b1-b40;
array p(*) p1-p40;
array u(*) u1-u40;
array z(*) z1-z40;

     call STREAMINIT(&seed);
       do j = 1 to 1400;
          theta = rannor(0); *theta for ref group;
              do i = 1 to 40;
               z(i) = 1.702*a(i)*(theta-b(i)); *d=1.702;
                     p(i) = exp(z(i))/(1+exp(z(i)));
                     u(i) = RAND("Uniform");
                   if u(i)<=p(i) then resp(i)=1; else resp(i)=0;
            end;
                  ExpScore_sim=sum(of p(*));
                  output;
       end;
run;

data respU1LC2;
set item_para2;

array resp(*) resp1-resp40;
array a(*) a1-a40;
array b(*) b1-b40;
array p(*) p1-p40;
array u(*) u1-u40;
array z(*) z1-z40;
```

```
     call STREAMINIT(&seed);
       do j = 1 to 600;
          theta = rannor(0); *theta for focal group;
              do i = 1 to 40;
               z(i) = 1.702*a(i)*(theta-b(i)); *d=1.702;
                     p(i) = exp(z(i))/(1+exp(z(i)));
                     u(i) = RAND("Uniform");
                   if u(i)<=p(i) then resp(i)=1; else resp(i)=0;
            end;
                  ExpScore_sim=sum(of p(*));
                  output;
       end;
run;

/* merge to a single dataset */
data respU1LC1_new;
set respU1LC1 (keep=resp1-resp40 theta ExpScore_sim);LC=0; run;

data respU1LC2_new;
set respU1LC2 (keep=resp1-resp40 theta ExpScore_sim);LC=1; run;

data respU1;
merge respU1LC1_new respU1LC2_new;
by LC;
id=_n_;
random=ranuni(27407349);
run;

proc sort data=respU1;
by LC random;
run;

/* assign manifest membership */
data respU1_cov;
set respU1;
id_ran=_n_;
if id_ran le 400+&x then do;
      cov=0; end;
if 400+&x < id_ran <= 1400 then do;
      cov=1; end;
if 1400 < id_ran <= 2000-&x then do;
      cov=0; end;
if id_ran > 2000-&x then do;
    cov=1; end;
drop random id_ran;
run;
proc sort data=respU1_cov;
by id;
run;

data _null_;
set respU1_cov;
file "&dir\resp.dat" DLM=','; /* data for Mplus */
put resp1-resp40 theta LC id cov;
run;

x 'mplus F:\code\mle_Cov.inp mplus F:\code\mle_Cov.out';
```

```
 data class; /*read in LC membership*/
 infile "&dir\class.dat";
 input y1-y40 COV F C_F CPROB1 CPROB2 C;
 id=_n_;
 if c=1 then LC_est=0;
 if c=2 then LC_est=1;
 OR=&OR;
 rep=&ii;
 run;

%include 'F:\code\dif.sas';run; /*compute DIF*/

data para; /*read in para*/
infile "&dir\para.dat";
input LC1Lamda1-LC1Lamda40 LC2Lamda1-LC2Lamda40 LC1tau1-LC1tau40 LC2tau1-
LC2tau40 C1Inter C1Cov DIFb1-DIFb40 DIFa1-DIFa40 LC1LamdaSE1-LC1LamdaSE40
LC2LamdaSE1-LC2LamdaSE40 LC1tauSE1-LC1tauSE40 LC2tauSE1-LC2tauSE40 C1InterSE
C1CovSE DIFbSE1-DIFbSE40 DIFaSE1-DIFaSE40 H0 H0Scaling NumOfPara AIC BIC ABIC
Entropy CondNum;
OR=&OR;
rep=&ii;
diftype=0;difnum=0;difeffect=0;impact=0;cov=1;est=0;
run;

%include 'F:\code\RMSE.sas';run; /*compute RMSE and Acc*/


data final;
merge para RMSE power2 CR2;
ccnum=powernum+crnum;
type1=1-CR; type2=1-power; cc=ccnum/40;
run;

proc append base=Cov data=final;
run;

 %put "Ending Loop &ii";
      %end;
%end;
%mend;

data; options noxwait;
%datasim;run;
```

APPENDIX C

MLE IN MPLUS

```
TITLE: Mixture 2PL without cov for no impact groups
DATA: FILE IS resp.dat;
VARIABLE: NAMES ARE y1-y40 theta LC id cov;
USEVARIABLES ARE y1-y40;
CATEGORICAL = y1-y40;
CLASSES = c(2);
ANALYSIS: TYPE = MIXTURE;
STARTS = 0;
ESTIMATOR = MLR;
ALGORITHM=INTEGRATION;
MODEL:
%OVERALL%
f by y1-y40;
f@1;
%c#1% !For Latent Group 1
f BY
y1*0.83962 (a11)
y2*1.16344 (a12)
y3*1.42861 (a13)
y4*0.51953 (a14)
y5*0.71311 (a15)
y6*1.14013 (a16)
y7*0.82456 (a17)
y8*0.78848 (a18)
y9*1.16716 (a19)
y10*1.33606 (a110)
y11*0.44606 (a111)
y12*0.67009 (a112)
y13*0.99443 (a113)
y14*0.68752 (a114)
y15*1.35674 (a115)
y16*0.66216 (a116)
y17*0.45865 (a117)
y18*1.78664 (a118)
y19*2.03085 (a119)
y20*1.73668 (a120)
y21*1.57925 (a121)
y22*0.54061 (a122)
y23*1.97391 (a123)
y24*0.65948 (a124)
y25*0.7942 (a125)
y26*1.18033 (a126)
y27*1.63773 (a127)
```

```
y28*1.07664 (a128)
y29*1.05928 (a129)
y30*1.36463 (a130)
y31*1.04468 (a131)
y32*2.78423 (a132)
y33*0.64483 (a133)
y34*1.29112 (a134)
y35*0.92831 (a135)
y36*0.9682 (a136)
y37*1.05021 (a137)
y38*0.83922 (a138)
y39*1.16335 (a139)
y40*1.24204 (a140);

[y1$1*0.76805] (b11);
[y2$1*-2.06338] (b12);
[y3$1*-0.73567] (b13);
[y4$1*0.72773] (b14);
[y5$1*0.77243] (b15);
[y6$1*0.98478] (b16);
[y7$1*1.27472] (b17);
[y8$1*-0.21978] (b18);
[y9$1*1.81587] (b19);
[y10$1*-0.49043] (b110);
[y11$1*0.40092] (b111);
[y12$1*-0.00508] (b112);
[y13$1*-0.75655] (b113);
[y14$1*0.77197] (b114);
[y15$1*-1.79896] (b115);
[y16$1*0.17812] (b116);
[y17$1*0.25901] (b117);
[y18$1*0.4481] (b118);
[y19$1*-0.70936] (b119);
[y20$1*-0.37609] (b120);
[y21$1*-0.92557] (b121);
[y22$1*-0.47226] (b122);
[y23$1*-3.4546] (b123);
[y24$1*-0.34418] (b124);
[y25$1*0.11247] (b125);
[y26$1*0.57599] (b126);
[y27$1*-1.7095] (b127);
[y28$1*-0.89621] (b128);
[y29$1*-1.66289] (b129);
[y30$1*-1.29466] (b130);
[y31$1*-0.53315] (b131);
[y32$1*1.32251] (b132);
[y33$1*0.52633] (b133);
[y34$1*1.55161] (b134);
[y35$1*-0.62573] (b135);
[y36$1*-1.18436] (b136);
[y37$1*-0.31385] (b137);
[y38$1*0.17377] (b138);
[y39$1*0.67429] (b139);
[y40$1*1.27261] (b140);
[f@0]; ! fix factor mean to zero

%c#2% !For Latent Group 2
```

```
f BY
y1*0.83962 (a21)
y2*1.16344 (a22)
y3*1.42861 (a23)
y4*0.51953 (a24)
y5*0.71311 (a25)
y6*1.14013 (a26)
y7*0.82456 (a27)
y8*0.78848 (a28)
y9*1.16716 (a29)
y10*1.33606 (a210)
y11*0.44606 (a211)
y12*0.67009 (a212)
y13*0.99443 (a213)
y14*0.68752 (a214)
y15*1.35674 (a215)
y16*0.66216 (a216)
y17*0.45865 (a217)
y18*1.78664 (a218)
y19*2.03085 (a219)
y20*1.73668 (a220)
y21*1.57925 (a221)
y22*0.54061 (a222)
y23*1.97391 (a223)
y24*0.65948 (a224)
y25*0.7942 (a225)
y26*1.18033 (a226)
y27*1.63773 (a227)
y28*1.07664 (a228)
y29*1.05928 (a229)
y30*1.36463 (a230)
y31*1.04468 (a231)
y32*2.78423 (a232)
y33*0.64483 (a233)
y34*1.29112 (a234)
y35*0.92831 (a235)
y36*0.9682 (a236)
y37*1.05021 (a237)
y38*0.83922 (a238)
y39*1.16335 (a239)
y40*1.24204 (a240);

[y1$1*0.76805] (b21);
[y2$1*-2.06338] (b22);
[y3$1*-0.73567] (b23);
[y4$1*0.72773] (b24);
[y5$1*0.77243] (b25);
[y6$1*0.98478] (b26);
[y7$1*1.27472] (b27);
[y8$1*-0.21978] (b28);
[y9$1*1.81587] (b29);
[y10$1*-0.49043] (b210);
[y11$1*0.40092] (b211);
[y12$1*-0.00508] (b212);
[y13$1*-0.75655] (b213);
[y14$1*0.77197] (b214);
[y15$1*-1.79896] (b215);
```

```
[y16$1*0.17812] (b216);
[y17$1*0.25901] (b217);
[y18$1*0.4481] (b218);
[y19$1*-0.70936] (b219);
[y20$1*-0.37609] (b220);
[y21$1*-0.92557] (b221);
[y22$1*-0.47226] (b222);
[y23$1*-3.4546] (b223);
[y24$1*-0.34418] (b224);
[y25$1*0.11247] (b225);
[y26$1*0.57599] (b226);
[y27$1*-1.7095] (b227);
[y28$1*-0.89621] (b228);
[y29$1*-1.66289] (b229);
[y30$1*-1.29466] (b230);
[y31$1*-0.53315] (b231);
[y32$1*1.32251] (b232);
[y33$1*0.52633] (b233);
[y34$1*1.55161] (b234);
[y35$1*-0.16158] (b235);
[y36$1*-0.70026] (b236);
[y37$1*0.21126] (b237);
[y38$1*0.59338] (b238);
[y39$1*1.25597] (b239);
[y40$1*1.89363] (b240);
[f@0]; ! fix factor mean to zero

OUTPUT: tech1 tech8 tech10;
SAVEDATA: FILE IS class.dat; RESULTS ARE para.dat;
SAVE=CPROBABILITIES fscores;
```

APPENDIX D

BAYESIAN ESTIMATION IN MPLUS

```
TITLE: Mixture 2PL without cov for non-impact groups
DATA: FILE IS resp.dat;
VARIABLE: NAMES ARE y1-y40 theta LC id cov;
USEVARIABLES ARE y1-y40;
CATEGORICAL = y1-y40;
CLASSES = c(2);
ANALYSIS: TYPE = MIXTURE;
ESTIMATOR = BAYES;
CHAINS=2;
STARTS=0; ! do ML from assigned starts
STVALUES=ML; ! start ML; ! start Bayes from the best ML solution
POINT=MODE;
BITERATIONS = 20000;
BCONVERGENCE=.05
THIN=50;
PROCESSOR = 2;
MODEL:
%OVERALL%
f by y1-y40;
f@1;
%c#1% !For Latent Group 1
f BY
y1*0.83962 (a1)
y2*1.16344 (a2)
y3*1.42861 (a3)
y4*0.51953 (a4)
y5*0.71311 (a5)
y6*1.14013 (a6)
y7*0.82456 (a7)
y8*0.78848 (a8)
y9*1.16716 (a9)
y10*1.33606 (a10)
y11*0.44606 (a11)
y12*0.67009 (a12)
y13*0.99443 (a13)
y14*0.68752 (a14)
y15*1.35674 (a15)
y16*0.66216 (a16)
y17*0.45865 (a17)
y18*1.78664 (a18)
y19*2.03085 (a19)
y20*1.73668 (a20)
y21*1.57925 (a21)
```

```
y22*0.54061 (a22)
y23*1.97391 (a23)
y24*0.65948 (a24)
y25*0.7942 (a25)
y26*1.18033 (a26)
y27*1.63773 (a27)
y28*1.07664 (a28)
y29*1.05928 (a29)
y30*1.36463 (a30)
y31*1.04468 (a31)
y32*2.78423 (a32)
y33*0.64483 (a33)
y34*1.29112 (a34)
y35*0.92831 (a35)
y36*0.9682 (a36)
y37*1.05021 (a37)
y38*0.83922 (a38)
y39*1.16335 (a39)
y40*1.24204 (a40);

[y1$1*0.76805] (b1);
[y2$1*-2.06338] (b2);
[y3$1*-0.73567] (b3);
[y4$1*0.72773] (b4);
[y5$1*0.77243] (b5);
[y6$1*0.98478] (b6);
[y7$1*1.27472] (b7);
[y8$1*-0.21978] (b8);
[y9$1*1.81587] (b9);
[y10$1*-0.49043] (b10);
[y11$1*0.40092] (b11);
[y12$1*-0.00508] (b12);
[y13$1*-0.75655] (b13);
[y14$1*0.77197] (b14);
[y15$1*-1.79896] (b15);
[y16$1*0.17812] (b16);
[y17$1*0.25901] (b17);
[y18$1*0.4481] (b18);
[y19$1*-0.70936] (b19);
[y20$1*-0.37609] (b20);
[y21$1*-0.92557] (b21);
[y22$1*-0.47226] (b22);
[y23$1*-3.4546] (b23);
[y24$1*-0.34418] (b24);
[y25$1*0.11247] (b25);
[y26$1*0.57599] (b26);
[y27$1*-1.7095] (b27);
[y28$1*-0.89621] (b28);
[y29$1*-1.66289] (b29);
[y30$1*-1.29466] (b30);
[y31$1*-0.53315] (b31);
[y32$1*1.32251] (b32);
[y33$1*0.52633] (b33);
[y34$1*1.55161] (b34);
[y35$1*-0.62573] (b35);
[y36$1*-1.18436] (b36);
[y37$1*-0.31385] (b37);
```

```
[y38$1*0.17377] (b38);
[y39$1*0.67429] (b39);
[y40$1*1.27261] (b40);
[f@0]; ! fix factor mean to zero

%c#2% !For Latent Group 2
f BY
y1*0.83962 (c1)
y2*1.16344 (c2)
y3*1.42861 (c3)
y4*0.51953 (c4)
y5*0.71311 (c5)
y6*1.14013 (c6)
y7*0.82456 (c7)
y8*0.78848 (c8)
y9*1.16716 (c9)
y10*1.33606 (c10)
y11*0.44606 (c11)
y12*0.67009 (c12)
y13*0.99443 (c13)
y14*0.68752 (c14)
y15*1.35674 (c15)
y16*0.66216 (c16)
y17*0.45865 (c17)
y18*1.78664 (c18)
y19*2.03085 (c19)
y20*1.73668 (c20)
y21*1.57925 (c21)
y22*0.54061 (c22)
y23*1.97391 (c23)
y24*0.65948 (c24)
y25*0.7942 (c25)
y26*1.18033 (c26)
y27*1.63773 (c27)
y28*1.07664 (c28)
y29*1.05928 (c29)
y30*1.36463 (c30)
y31*1.04468 (c31)
y32*2.78423 (c32)
y33*0.64483 (c33)
y34*1.29112 (c34)
y35*0.92831 (c35)
y36*0.9682 (c36)
y37*1.05021 (c37)
y38*0.83922 (c38)
y39*1.16335 (c39)
y40*1.24204 (c40);

[y1$1*0.76805] (d1);
[y2$1*-2.06338] (d2);
[y3$1*-0.73567] (d3);
[y4$1*0.72773] (d4);
[y5$1*0.77243] (d5);
[y6$1*0.98478] (d6);
[y7$1*1.27472] (d7);
[y8$1*-0.21978] (d8);
[y9$1*1.81587] (d9);
```

```
[y10$1*-0.49043] (d10);
[y11$1*0.40092] (d11);
[y12$1*-0.00508] (d12);
[y13$1*-0.75655] (d13);
[y14$1*0.77197] (d14);
[y15$1*-1.79896] (d15);
[y16$1*0.17812] (d16);
[y17$1*0.25901] (d17);
[y18$1*0.4481] (d18);
[y19$1*-0.70936] (d19);
[y20$1*-0.37609] (d20);
[y21$1*-0.92557] (d21);
[y22$1*-0.47226] (d22);
[y23$1*-3.4546] (d23);
[y24$1*-0.34418] (d24);
[y25$1*0.11247] (d25);
[y26$1*0.57599] (d26);
[y27$1*-1.7095] (d27);
[y28$1*-0.89621] (d28);
[y29$1*-1.66289] (d29);
[y30$1*-1.29466] (d30);
[y31$1*-0.53315] (d31);
[y32$1*1.32251] (d32);
[y33$1*0.52633] (d33);
[y34$1*1.55161] (d34);
[y35$1*-0.16158] (d35);
[y36$1*-0.70026] (d36);
[y37$1*0.21126] (d37);
[y38$1*0.59338] (d38);
[y39$1*1.25597] (d39);
[y40$1*1.89363] (d40);
[f@0]; ! fix factor mean to zero
MODEL PRIORS:
a1-a40 ~ N(0,1);
b1-b40 ~ N(0,1);
c1-c40 ~ N(0,1);
d1-d40 ~ N(0,1);
OUTPUT: tech8;
PLOT: TYPE=PLOT2;
SAVEDATA: FILE IS class_bayes.dat; BPARAMETERS=para_bayes.dat;
SAVE=fscores(100);
```

BIBLIOGRAPHY

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. Journal of educational and behavioral Statistics, 22(1), 47-76.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. Journal of Educational and Behavioral Statistics, 17(3), 251-269.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. Journal of the American statistical Association, 88(422), 669-679.

Aryadoust, V. (2015). Fitting a mixture Rasch model to English as a foreign language listening tests: The role of cognitive and background variables in explaining latent differential item functioning. International Journal of Testing, 15(3), 216-238.

Aryadoust, V., & Zhang, L. (2016). Fitting the mixed Rasch model to a reading comprehension test: Exploring individual difference profiles in L2 reading. Language Testing, 33(4), 529-553.

Asparouhov, T., & Muthén, B. (2010). Bayesian analysis using Mplus: Technical implementation. Manuscript submitted for publication.

Asparouhov, T., & Muthén, B. (2016). IRT in Mplus. Version 2. Technical report. Retrieved from, https://www. statmodel. com/download/MplusIRT. pdf.

Baker, F. B., & Kim, S. H. (Eds.). (2004). Item response theory: Parameter estimation techniques. CRC Press.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. Journal of Educational and Behavioral Statistics, 26(4), 381-409.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. Journal of Educational Measurement, 39(4), 331-348.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. Behaviormetrika, 23(1), 67-95.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. Psychometrika, 64(2), 153-168.

Brooks, S. (1998). Markov chain Monte Carlo method and its application.Journal of the royal statistical society: series D (the Statistician), 47(1), 69-100.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics, 7(4), 434-455.

Brooks, S., & Roberts, G. O. (1998). Convergence assessments of Markov chain Monte Carlo algorithms. Statistics and Computing, 8, 319-335.

Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. Journal of the Royal Statistical Society. Series B (Methodological), 473-484.

Chen, Y. F., & Jiao, H. (2014). Exploring the utility of background and cognitive variables in explaining latent differential item functioning: An example of the PISA 2009 reading assessment. Educational Assessment, 19(2), 77-96.

Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. The American statistician, 49(4), 327-335.

Cho, Y. (2013). The mixture distribution polytomous Rasch model used to account for response styles on rating scales: A simulation study of parameter recovery and classification accuracy. Unpublished doctoral dissertation, University of Maryland, College Park, Maryland.

Cho, S. J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. Journal of Educational and Behavioral Statistics, 35(3), 336-370.

Cho, S. J., Cohen, A. S., & Kim, S. H. (2006). An investigation of priors on the probabilities of mixtures in the mixture Rasch model. In International Meeting of the Psychometric Society: The 71st annual meeting of the Psychometric Society, Montreal, Canada.

Cho, S. J., Cohen, A. S., & Kim, S. H. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. Journal of Statistical Computation and Simulation, 83(2), 278-306.

Cho, S. J., Cohen, A. S., & Kim, S. H. (2014). A mixture group bifactor model for binary responses. Structural Equation Modeling: A Multidisciplinary Journal, 21(3), 375-395.

Cho, S. J., Cohen, A. S., Kim, S. H., & Bottge, B. (2010). Latent transition analysis with a mixture item response theory measurement model. Applied Psychological Measurement, 34(7), 483-504.

Cho, S. J., Suh, Y., & Lee, W. Y. (2015). An NCME Instructional Module on Latent DIF Analysis Using Mixture Item Response Models. Educational Measurement: Issues and Practice.

Choi, Y. J. (2014). Metric Identification in Mixture IRT Models. uga.

Choi, H. J. (2010). A model that combines diagnostic classification assessment with mixture item response theory models. Unpublished doctoral dissertation, University of Georgia, Athens, Georgia.

Choi, Y., Alexeev, N., & Cohen, A. (2014). DIF Analysis using a Mixture 3PL Model with a Covariate on the TIMSS 2007 Mathematics Test. In KAERA Research Forum (Vol. 1, No. 1, pp. 4-14).

Chung, H., Flaherty, B. P., & Schafer, J. L. (2006). Latent class logistic regression: application to marijuana use and attitudes among high school seniors. Journal of the Royal Statistical Society: Series A (Statistics in Society), 169(4), 723-743.

Clogg, C. C. (1988). Latent class models of measuring. In R. Langeheine & J. Rost (Eds.), Latent trait and latent class models (pp. 173-206). New York: Plenum.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. Journal of Educational Measurement, 42(2), 133-148.

Cohen, A. S., Wollack, J. A., Bolt, D. M., & Mroch, A. A. (2002). A mixture Rasch model analysis of test speededness. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association, 91(434), 883-904.

Cox, P. (2000). Regional and gender differences in mathematics achievement. Journal of Research in Rural Education, 16(1), 22-29.

Dai, Y. (2009). A mixture Rasch model with a covariate: a simulation study via Bayesian Markov Chain Monte Carlo Estimation. Unpublished doctoral dissertation, University of Maryland, College Park, Maryland.

Dai, Y. (2013). A mxture Rasch model with a covariate: A simulation study via Bayesian Markov Chain Monte Carlo estimation. Applied Psychological Measurement, 0146621612475076.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. Journal of the American Statistical Association, 83(401), 173-178.

De Ayala, R. J. (2013). The theory and practice of item response theory. Guilford Publications.

De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. International Journal of Testing, 2(3-4), 243-276.

De Ayala, R. J., & Santiago, S. Y. (2017). An introduction to mixture item response theory models. Journal of School Psychology, 60, 25-40.

DeMars, C. E. (2000). Test stakes and item format interactions. Applied Measurement in Education, 13(1), 55-77.

DeMars, C. E., & Lau, A. (2011). Differential Item Functioning Detection With Latent Classes: How Accurately Can We Detect Who Is Responding Differentially?. Educational and Psychological Measurement, 71(4), 597-616.

Driana, E. (2007). Gender differential item functioning on a ninth-grade mathematics proficiency test in appalachian ohio (Doctoral dissertation, Ohio University).

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. Journal of the american statistical association,90(430), 577-588.

Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. Journal of Modern Applied Statistical Methods, 11(1), 14.

Formann, A. K. (1985). Constrained latent class models: Theory and applications. British Journal of Mathematical and Statistical Psychology, 38, 87-111.

Formann, A. K. (1989). Constrained latent class models: Some further applications. British Journal of Mathematical and Statistical Psychology, 42, 37-54.

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

Fox, J. P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika, 66(2), 271-288.

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. Journal of Educational Measurement, 47(4), 432-457.

Garrett, E. S., & Zeger, S. L. (2000). Latent class model diagnosis. Biometrics, 56(4), 1055-1067.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. Journal of the American statistical association, 85(410), 398-409.

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. Handbook of Markov chain Monte Carlo, 163-174.

Gill, J. (2014). Bayesian methods: A social and behavioral sciences approach (Vol. 20). CRC press.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Review of Educational Research, 42, 237-288.

Gnaldi, M., Bacci, S., & Bartolucci, F. (2016). A multilevel finite mixture item response model to cluster examinees and schools. Advances in Data Analysis and Classification, 10(1), 53-70.

Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. Journal of Educational Measurement, 46(3), 314-329.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2(4), 313-334.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications (Vol. 7). Springer Science & Business Media.

Hoijtink, H. (1998). Constrained latent class analysis using the Gibbs sampler and posterior predictive p-values: Applications to educational testing.Statistica Sinica, 691-711.

Hu, J., Leite, W. L., & Gao, M. (2017). An evaluation of the use of covariates to assist in class enumeration in linear growth mixture modeling. Behavior Research Methods, 1-12.

Huang, D., Brecht, M. L., Hara, M., & Hser, Y. I. (2010). Influences of covariates on growth mixture modeling. Journal of drug issues, 40(1), 173-194.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. Journal of Educational and Behavioral Statistics, 25(3), 285-306.

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Statistical Science, 50-67.

Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response models in R. Journal of Statistical Software, 20(10), 1-24.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. Journal of Educational Measurement, 27(4), 307-327.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., & Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Statistics in medicine, 24(15), 2401-2428.

Lee, J., & Mcintire, W. G. (2000). Interstate variation in the mathematics achievement of rural and nonrural students. Journal of Research in Rural Education, 16(3), 168-181.

Levy, R. (2009). The rise of Markov chain Monte Carlo estimation for psychometric modeling. Journal of Probability and Statistics, 2009.

Li, F., Cohen, A., Bottge, B., & Templin, J. (2015). A Latent Transition Analysis Model for Assessing Change in Cognitive Skills. Educational and Psychological Measurement. 76(2), 181-204.

Li, F., Cohen, A. S., Kim, S. H., & Cho, S. J. (2009). Model selection methods for mixture dichotomous IRT models. Applied Psychological Measurement.

Li, L., & Hser, Y. I. (2011). On inclusion of covariates for class enumeration of growth mixture models. Multivariate behavioral research, 46(2), 266-302.

Li, T., Jiao, H., & Macready, G. B. (2015). Different Approaches to Covariate Inclusion in the Mixture Rasch Model. Unpublished doctoral dissertation, University of Maryland.

Li, T., Jiao, H., & Macready, G. B. (2016). Different approaches to covariate inclusion in the mixture Rasch model. Educational and Psychological Measurement, 76(5), 848-872.

Linn, M. C., & Hyde, J. S. (1989). Gender, mathematics, and science. Educational Researcher, 18(8), 17-27.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters upon DIF detection using the free-baseline likelihood ratio test. Applied Psychological Measurement, 33, 251-265.

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

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley.

Lu, R., & Jiao, H. (2009, April). Detecting DIF using the mixture Rasch model. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Lubke, G., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. Structural Equation Modeling, 14(1), 26-47.

Lynch, S. M. (2007). Introduction to applied Bayesian statistics and estimation for social scientists. Springer Science & Business Media.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. Multivariate Behavioral Research, 45(6), 975-999.

Merkle, E. C. (2005). Bayesian Estimation of Factor Analysis Models with Incomplete Data. Unpublished doctoral dissertation, Ohio State University.

Meyer, J. P. (2010). A mixture Rasch model with item response time components. Applied Psychological Measurement. 34(7) 521–538.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. Applied Psychological Measurement, 11(1), 81-91.

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. Psychometrika, 54(4), 661-679.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. Psychometrika, 55(2), 195-215.

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. Psychometrika, 61(1), 41-71.

Müller, U. K. (2012). Measuring prior sensitivity and prior informativeness in large Bayesian models. Journal of Monetary Economics, 59(6), 581-597.

Muthén B. O. (2004). Latent variable analysis: growth mixture modeling and related techniques for longitudinal data, in Handbook of Quantitative Methodology for the Social Sciences, ed Kaplan D., editor. (Newbury Park, CA: Sage), 345–368.

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

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. Applied Psychological Measurement, 18(4), 315-328.

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

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. Applied Psychological Measurement, 16(3), 237-248.

Paek, I., & Cho, S. J. (2015). A note on parameter estimate comparability: across latent classes in mixture IRT modeling. Applied Psychological Measurement, 39(2), 135-143.

Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. Journal of educational and behavioral statistics, 24(4), 342-366.

Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. Journal of educational and behavioral Statistics, 24(2), 146-178.

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.

Phillips, D. B. and Smith, A. F. M. (1996). Bayesian model comparison via jump diffusions. In Markov Chain Monte Carlo in Practice, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (eds), 215-239. London: Chapman and Hall.

Puhan, G., Moses, T. P., Yu, L., & Dorans, N. J. (2009). Using Log ‐ Linear Smoothing to Improve Small ‐ Sample DIF Estimation. Journal of Educational Measurement, 46(1), 59-83.

Raju, N.S., Bode, R.K., & Larsen, V.S. (1989). An empirical assessment of the Mantel-Haenszel statistic to detect differential item functioning. Applied Measurement in Education, 2, 1-13.

Randhawa, B. S., & Hunt, D. (1987). Sex and rural-urban differences in standardized achievement scores and mathematics subskills. Canadian Journal of Education/Revue canadienne de l'education, 137-151.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational and Behavioral Statistics, 4(3), 207-230.

Roberts, G. O., & Smith, A. F. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. Stochastic processes and their applications, 49(2), 207-216.

Roscigno, V. J., Tomaskovic-Devey, D., & Crowley, M. L. (2006). Education and the inequalities of place. Social Forces, 84(4), 2121-2145.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. Applied Psychological Measurement, 14(3), 271-282.

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

Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In I. W. Moelenaar (Ed.), Rasch models: Foundations, recent developments and applications (pp. 257-268). New York, NY: Springer Verlag.

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

Samuelsen, K. M. (2005). Examining differential item functioning from a latent class perspective (Doctoral dissertation). Available from ProQuest Dissertations and Theses database.

Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. Child development, 85(3), 842-860.

Sen, S., Cohen, A. S., & Kim, S. H. (2014, November). Robustness of mixture IRT models to violations of latent normality. In Quantitative Psychology Research: The 78th Annual Meeting of the Psychometric Society (Vol. 89, p. 27). Springer.

Smit, A., Kelderman, H., & van der Flier, H. (1999). Collateral information and mixed Rasch models. Methods of Psychological Research Online, 4(3), 19-32.

Smit, A., Kelderman, H., & van der Flier, H. (2000). The mixed Birnbaum model: Estimation using collateral information. Methods of Psychological Research Online, 5(4), 31-43.

Smit, A., Kelderman, H., & van der Flier, H. (2003). Latent trait latent class analysis of an eysenck personality questionnaire. Methods of Psychological Research Online, 8 (3), 23-50.

Svetina, D., & Rutkowski, L. (2014). Detecting differential item functioning using generalized logistic regression in the context of large-scale assessments. Large-scale Assessments in Education, 2(1), 4.

Tay, L., Huang, Q., & Vermunt, J. K. (2016). Item Response Theory With Covariates (IRT-C) Assessing Item Recovery and Differential Item Functioning for the Three-Parameter Logistic Model. Educational and Psychological Measurement, 76(1), 22-42.

Tay, L., Newman, D. A., & Vermunt, J. K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. Organizational Research Methods, 14(1), 147-176.

Tay, L., Vermunt, J. K., & Wang, C. (2013). Assessing the item response theory with covariate (IRT-C) procedure for ascertaining differential item functioning. International Journal of Testing, 13(3), 201-222.

Turner, R. M., Omar, R. Z., & Thompson, S. G. (2001). Bayesian methods of analysis for cluster randomized trials with binary outcome data. Statistics in medicine, 20(3), 453-472.

Van der Heijden, P. G., Dessens, J., & Bockenholt, U. (1996). Estimating the concomitant-variable latent-class model with the EM algorithm. Journal of Educational and Behavioral Statistics, 21(3), 215-229.

Van Nijlen, D., & Janssen, R. (2008). Mixture IRT-models as a means of DIF-detection: Modelling spelling in different grades of primary school. Paper presented at the annual meeting of the National Council of Measurement in Education, New York, NY.

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. Statistical Methods in Medical Research, 17(1), 33-51.

Vermunt, J. K. (2010) Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. Political Analysis, 18, 450-469.

von Davier, M. (2005). mdltm: Software for the general diagnostic model and for estimating mixture of multidimensional discrete latent traits models [Computer software]. Princeton, NJ: Educational Testing Service.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In I. W. Moelenaar (Ed.), Rasch models: Foundations, recent developments and applications (pp. 371-379). New York, NY: Springer Verlag.

von Davier, M., & Rost, J. (2006). 19 Mixture Distribution Item Response Models. Handbook of statistics, 26, 643-661.

Von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. Applied Psychological Measurement, 28(6), 389-406.

Wang, A. (2011). A Mixture Cross-classification IRT Model for Test Speededness. Unpublished doctoral dissertation, University of Georgia, Athens, Georgia.

Wang, W.-C. (2004). Effects of anchor item methods on differential item functioning detection within the family of Rasch models. The Journal of Experimental Education, 72, 221-261.

Willse, J. T. (2010). Mixture Rasch models with joint maximum likelihood estimation. Educational and Psychological Measurement, 17(1), 5-19.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation.Applied psychological measurement, 26(3), 339-352.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. Journal of Educational Measurement, 40(4), 307-330.

Wurpts, I. C., & Geiser, C. (2014). Is adding more indicators to a latent class analysis beneficial or detrimental? Results of a Monte-Carlo study. Frontiers in psychology, 5, 920.

Zhu, X. (2013). Distinguishing continuous and discrete approaches to multilevel mixture IRT models: A model comparison perspective. Unpublished doctoral dissertation, University of Maryland, College Park, Maryland.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. Journal of Educational Measurement, 26(1), 55-66.