ABSTRACT

Title of Document: DISTINGUISHING CONTINUOUS AND DISCRETE APPROACHES TO MULTILEVEL MIXTURE IRT MODELS: A MODEL COMPARISON PERSPECTIVE

Xiaoshu Zhu, Doctor of Philosophy, 2013

Directed By: Dr. Robert W. Lissitz
Department of Human Development and Quantitative Methodology

The current study introduced a general modeling framework, multilevel mixture IRT (MMIRT) which detects and describes characteristics of population heterogeneity, while accommodating the hierarchical data structure. In addition to introducing both continuous and discrete approaches to MMIRT, the main focus of the current study was to distinguish continuous and discrete MMIRT models from a model comparison perspective. A simulation study was conducted to evaluate the impact of class separation, cluster size, proportion of mixture, and between-group ability variance on the model performance of a set of MMIRT models. The behavior of information-based fit criteria in distinguishing between discrete and continuous MMIRT models was also investigated. An empirical analysis was presented to illustrate the application of MMIRT models.

Results suggested that class separation, and between-group ability variance had significant impact on MMIRT model performance. Discrete MMIRT models with fewer group-level latent classes performed consistently better on parameter and classification recovery than the continuous MMIRT model and the discrete models with more latent classes at the group level. Despite the poor performance of the continuous MMIRT model, it was favored over the discrete models by most fit indices. The AIC, AIC3, AICC, and the modifications of AIC and ssBIC were more sensitive to the discreteness in random effect distribution, compared to the CAIC, BIC, their modifications, and ssBIC. The latter ones had a higher tendency to select continuous MMIRT model as the best fitting model, regardless of the true distribution of random effects.

DISTINGUISHING CONTINUOUS AND DISCRETE APPROACHES TO MULTILEVEL MIXTURE IRT MODELS: A MODEL COMPARISON PERSPECTIVE

By

Xiaoshu Zhu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Professor Robert W. Lissitz, Chair
Professor Robert G. Croninger
Professor Jeffrey R. Harring
Professor Hong Jiao
Professor George B. Macready

Dedication

For my beloved parents,

Kaiyu Zhu and Xiaoxing Zhang,

who give me endless love and support.

Acknowledgements

I would first like to express my deep gratitude to my advisor, Dr. Lissitz for his support and guidance throughout my years in the doctoral program and especially, throughout the dissertation process. All these years, you are more like a close friend who I can share every joyful moment with. Thank you for always being there whenever I need help. Thank you for your encouragement whenever I lose confidence in myself. Without your insight and instruction, I could not complete this study.

I would also like to thank my committee members, Dr. Macready, Dr. Harring, Dr. Jiao and Dr. Croninger for their time and recommendations on my study. I enjoyed every chance to learn from you. Thank you for advising and inviting me to join your studies, where I gradually built up my own research skills. I also want to thank Dr. Mislevy, Dr. Hancock, and Dr. Rupp for their help and support during my stay in EDMS. I would also like to give special thanks to Dr. Stapleton for introducing me to Westat, where I will make the transition from a student to a professional. To my dear fellow students, without you the life in UMD cannot be colorful and joyful.

To whom this is dedicated, my dearest parents, for without them, these all would not be possible. Also to my dear husband, Ling Hung, for his love, company and understanding in the past four years, I smile because of you.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The No Child Left Behind (NCLB, 2001) Act and Race to the Top (2009) both require psychometricians to help educators evaluate schools and teachers (Lissitz, 2012). Since their enactment, complex psychometric models have been developed to connect student academic achievement with their teachers and their schools. The hierarchical nature of educational data can be represented appropriately in multilevel models (Bryk & Raudenbush, 1992; Goldstein, 2010) with students nested within group level units such as teachers and schools. The development of multilevel analyses is driving interest in identifying the characteristics of effective schools and teachers and the criteria for measuring effectiveness (Fox, 2005).

Two general trends exist to evaluate school and teacher effectiveness, either taking a longitudinal approach or focusing on measures at a single time point. While value-added models (VAMs; Ballou, Sanders, & Wright, 2004; Kane, Rockoff, & Staiger, 2006; Lissitz, 2005; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Sanders, Saxton, & Horn, 1997) estimate the contribution of teachers or schools to the achievement growth of students as they progress through grades, other multilevel models concentrate on investigating the impact of context effects (e.g., school and teacher effects) of student performance on achievement assessment. The models proposed in the current study are of the latter type.

1.1 Statement of Problem

An implicit assumption underlying the study of context effect is that teacher and school differences on effectiveness cause between-group variation of student

academic performance. When context effects are modeled as latent variables, one issue that draws great interest is whether such variables are better described as continuous or categorical.

Context effects have been modeled as either continuous or categorical latent variables in the existing literature. The contribution of context effect on student achievement is often judged in terms of the percentage of variance accounted for by the teacher and school levels (Rowan, Correnti, & Miller, 2002). Hierarchical linear models (HLMs) have been widely implemented to decompose the variance in student achievement into within- and between-group components. While conventional HLMs describe the overall contribution of school and teacher effectiveness, some later extensions of multilevel models such as cross-classified models (CCM, Raudenbush & Bryk, 2002) and layered models (Ballou et al., 2004; Sanders et al., 1997) separate the persistent contributions of past teachers to current test scores. In those models, the context effect is assumed to be a continuous variable. Meanwhile, context effect is modeled as a set of latent classes in other multilevel models to capture population heterogeneity at the group level. The presence of unobserved group-level subpopulations can partly explain student difference on academic performance. For example, the multilevel growth mixture model with between-group mixtures (Palardy & Vermunt, 2010) provides a means of classifying schools into homogeneous classes in terms of the properties of their student mean achievement growth trajectories.

The current study focuses on examining context effect reflected on item-level responses. More precisely, the question of interest is whether teaching practice affects the probability of students being clustered into a particular latent class that is

characterized by differential item functioning (DIF). DIF arises when the property of a particular item differs among examinees conditioning on their ability level. Recent studies have revealed that the differences in unobserved attributes, such as curricular experience may, in part, cause the DIF (Cohen, Gregg, & Deng, 2005). From a teaching practice perspective, this difference may reflect distinctive school and teacher effects on student learning. However, to assume that students are equally affected by their teachers and schools is unrealistic. It is widely accepted that a certain teaching practice will be effective with one type of students but not with others. Even given the same curricular practice, the perceived curricular experience can differ. Reflecting on item responses, DIF can exist among students from the same class or school. Thus, the investigation of context effect on DIF can provide valuable information regarding school or teacher effect.

Mixture modeling is a statistical tool for identifying latent groups of individuals (McLachlan & Peel, 2000). As applied to measurement models, mixture IRT models are gaining in popularity in investigating possible latent causes of DIF (Cohen & Bolt, 2005; Samuelsen, 2005). To investigate context effect on DIF, multilevel extensions of conventional mixture IRT models are developed to appropriately accommodate the hierarchical structure. Multilevel mixture IRT (MMIRT) models are of this kind, and can be derived from multilevel mixture generalized models proposed by Vermunt (2008a). Compared to conventional mixture models, multilevel mixture models utilize either continuous random effects (a continuous approach) or a set of latent classes (a discrete approach) at the group level to assess variation in model parameters across group units.

It should be noted that Vermunt (2003) called the discrete approach "nonparametric" as opposed to the "parametric" approach that makes strong assumptions about the distribution of random effects. However, the term "nonparametric" does not imply "distribution free" in that the normal distribution assumption in the "parametric" approach is replaced by the assumption of a multinomial distribution. To avoid confusion, the current study uses "continuous" and "discrete", instead of "parametric" and "nonparametric", to describe the two approaches.

Whereas latent classes can suggest substantive group heterogeneity; an alternative hypothesis is that the identified classes represent simple variation on a continuum of a latent structure (Van Horn, et al., 2008). In other words, latent classes not only capture multidimensionality in latent structure, but also represent discreteness in a latent distribution (Markon & Krueger, 2006). Under certain circumstance, the distinction between continuous and discrete specifications of multilevel mixture models pertains to the presumption of latent distribution. For some studies that have vague theoretical hypotheses regarding distribution of group-level variation, it is rather reasonable to compare the continuous and discrete approaches in an exploratory manner. A model comparison perspective, thus, can be utilized to accomplish this goal.

1.2 Significance of the Study

The MMIRT framework offers practitioners an alternative solution to investigating context effects on item-level responses where both population heterogeneity and hierarchical structure are acknowledged. Models within the

MMIRT framework can be divided into two general categories depending on whether variation at the group-level is modeled as a continuous random variable or a set of discrete latent classes.

Discrete specifications of MMIRT models are not new. Cho (2007) and Vermunt (2008b) individually proposed two MMIRT models that are particularly utilized to identify school-level differences on item functioning while accommodating the hierarchical structure. Up to date, the continuous approach to MMIRT models is only a theoretical possibility. Instead, a similar modeling approach, the multilevel latent class analyses (MLCA) with continuous group-level random effects, has shown its potential to study the intervention effects in group randomized trials (Van Horn et al., 2008) and adolescent smoking typologies across communities (Henry & Muthén, 2010). An empirical study, then, is necessary to illustrate the specification of continuous MMIRT models and their implementation in practice.

Due to the complexity and large number of parameters, more often constraints are imposed on multilevel mixture models so that some parameters are not conditional on latent class membership. The decision with respect to constraints becomes even more complicated for models that introduce mixtures at both lower and higher levels. For instance, Asparouhov and Muthén (2008) described a multilevel mixture model where the model parameters differ across person-level latent classes but do not vary across group-level classes. Vermunt (2008b), in contrast, illustrated a similar model but with item parameters invariant among person-level classes. What constraints should be placed on the unrestricted model depends on the specific study purposes. In particular, if the main focus is to identify meaningful group

heterogeneity at lower-level while taking multilevel structure into account, model parameters may vary only between lower-level latent classes but remain constant across higher-level units. Even when latent classes are specified at higher level, they essentially represent variation among higher-level units instead of suggesting qualitative differences. In this scenario, the models with higher-level latent classes can be compared with the models using continuous random effects at the group level, leading to a test of discreteness versus continuousness.

Both the continuous and discrete approaches can be used to model the context effect as group-level random effects. The comparison between the two approaches shares a similar challenge with other latent variable models on how to use substantial evidence such as model fit criteria to support whether a continuous or a discrete specification more properly describes higher-level distributions.

Interest in methods of distinguishing between discrete and continuous latent distributions has grown in popularity in areas of clinical psychology and behavioral science. Such methods can also be applied to the comparison of the two approaches within the MMIRT framework. The key distinction between discrete and continuous latent variables is the number of values of latent distribution that further leads to non-negligible differences in fit and parameter estimates. The difference in fit provides important means for decision making about which latent structure, continuous or discrete, should be selected for a particular set of data. Previous studies limited their discussion to conventional latent variable models such as structural equation mixture modeling (Bauer & Curran, 2004), latent profile models (Lubke & Neale, 2006) and latent trait model (Markon & Krueger, 2006). As far as multilevel mixture models are

concerned, only one study (i.e., Henry & Muthén, 2010) has applied information-based model fit criteria to compare the continuous approach and the discrete approaches to MLCA. The BIC functioned so unstably that Henry and Muthén (2010) suggested more research to understand the performance of fit criteria in MLCA. Although information criteria have been widely used to select models with two distinctive types of latent variables, empirical studies are still required to fill in the blanks about the function of fit indices in multilevel mixture models.

1.3 The Purpose of the Study

The MMIRT framework is promising in that it allows the possibility to specify a variety of models with mixtures when data are hierarchical. Both continuous and discrete approaches to MMIRT are introduced, and special attention is given to MMIRT models with continuous random effects at the group level. In particular, the current study presents the connection between two possible ways of specifying group-level variation. The models illustrated in the current study are Rasch-model based and for dichotomously scored responses only.

The concern is to model the variation on probability of lower-level latent classes across higher-level units, hence, two restrictive MMIRT models are further proposed. These two types of models differ only with respect to the specifications of higher-level variation. Moreover, the question of whether model comparisons lead to correct model selection regarding the nature of group-level latent distributions, continuous or discrete, is explored with a simulation and an empirical application.

Although the framework is complex, few studies have been conducted to evaluate performance of MMIRT models in preparation for or in conjunction with

empirical analyses. The current study is the first attempt in the literature to use model fit criteria to distinguish between discrete and continuous MMIRT models. The purposes of this study are threefold: (1) to introduce two approaches to specify higher-level random effects in MMIRT, especially the continuous specification; (2) to investigate among various information criteria, which criterion works most effectively in identifying whether the latent distribution of random effects is continuous or discrete at higher level; and (3) to qualify the effect of class separation cluster size, proportion of mixture, and between-group ability variance on making this distinction.

1.4 Overview of Chapters

In the following chapter, the MMIRT framework is proposed after the introduction of a general latent variable modeling framework. Traditional mixture modeling approaches are extended to account for multilevel data structure. MMIRT models are special cases of the resultant multilevel mixture models.

In Chapter 2, the mixture IRT model, multilevel IRT models and multilevel latent class models, and how each of the model components is integrated into the MMIRT framework are discussed in detail. In particular, the mixture IRT model specifies the mixture proportion on person ability and item difficulty structure; the multilevel IRT model is included to identify ability variation at the group-level; and the multilevel latent class models contribute to the probability structure in MMIRT. The description focuses on why MMIRT models are promising approaches to represent complex data structure and identify heterogeneity at both lower and upper

levels. The incorporation of covariates from two levels in MMIRT is also addressed in this chapter.

Chapter 3 describes the technical issues with respect to the estimation methods and model selection. The latter part of Chapter 3 presents a simulation study designed to assess the power of model fit indices in distinguishing between the continuous and discrete specifications of MMIRT models.

The results of the simulation study are presented in Chapter 4, where the influence of manipulated factors on the recovery of model parameters and classification is presented first, followed by the discussion of how frequently the true models are selected using various model fit indices. In addition, the restrictive models are compared when applied to an empirical dataset sampled from the Maryland School Assessment (MSA). Chapter 5 summarizes the findings, and discusses potential limitations and future directions in the development of MMIRT models.

Chapter 2: Theoretical Background

MMIRT models proposed in this study are used explicitly for the detection of DIF while acknowledging the multilevel structure. In particular, the focus of the discussion of MMIRT is on how to model variation at the group-level. MMIRT models are special cases of multilevel mixture models (Vermunt, 2008a). Depending on the specification of latent variables, MMIRT models have two subtypes, a continuous approach with random effects following a continuous normal distribution and a discrete approach with a set of discrete latent classes. Both approaches are built upon the combination of mixture IRT models, multilevel IRT models as well as multilevel latent class models.

In this chapter, the general latent variable modeling framework is discussed first, followed by the introduction of three fundamental models of MMIRT.

2.1 Latent Variable Modeling Framework

Latent variables are defined as hypothetical constructs that can only be inferred from observed variables and are often differentiated in terms of their underlying distribution as continuous or categorical.

The nature of observed variables depends on the response format of the data, but the distinction between categorical and continuous latent variables is of considerable importance on a theoretical level (Lubke & Neale, 2006). A more common distinction between a categorical and continuous latent variable is the difference between a nominal (i.e., class, qualitative) latent variable that is necessarily categorical and discrete, and a metric (i.e., real numbers or interval) latent variable

that can be discrete or continuous. In this study, metric variables are assumed to be continuous, and the terms categorical and discrete are used interchangeably.

Conventional latent variable models with one type of latent variable can be classified into four general categories based on the types of observed and latent variables (Bartholomew & Knott, 1999), as shown in Table 2.1. Classical factor analysis (FA) is a general term for models characterized by continuous observed variables and continuous latent variables. When the observed variables are categorical, IRT models are obtained with continuous latent variables. The latent class analysis (LCA) deals with the situations when both observed and latent variables are categorical. This term and finite mixture model are used interchangeably in practice. If the categorical latent variables are inferred from continuous observed variables, a latent profile analysis (LPA) is obtained. All four analyses have been widely used in social and behavioral research.

Table 2.1 Classification of latent variable modeling

	Observed Variables	
Latent Variables	Continuous	Categorical
Continuous	Factor analysis	Item Response Theory
Categorical	Latent Profile analysis	Latent Class analysis

For the purpose of accommodating context effects, traditional latent variable models can be extended to include a higher level. Those models can be applied to the situations in which either a three-level univariate response or a two-level multivariate response data set are considered, where the former has an item or measurement level in addition to the person and group levels. Latent variables at the person level and

group level could be continuous (or random effects), discrete or a combination of these two. Thus, depending on the scale types of latent variable at the two levels, Vermunt (2007) proposed a nine-fold classification of latent variable models for multilevel data sets as shown in Table 2.2.

This classification is an expansion of the latent variable modeling framework introduced by Skrondal and Rabe-Hesketh (2004). This flexible framework provides a unifying theme of latent variables which can embrace various traditions such as growth modeling, multilevel modeling and finite mixture modeling. All categories except Category A1 (see Table 2.2) fall into a more general type labeled as multilevel mixture models; that is, models with latent classes at either one or at two levels (Vermunt, 2003, 2007). Compared with the traditional latent variable models, a multilevel mixture model contains either continuous random effects or a discrete latent variable at the group level to account for heterogeneity in model parameters across group units.

Table 2.2 Nine-fold classification of latent variable models for multilevel data sets

Person-level latent variables	Group-level latent variables		
	Continuous	Categorical	Combination
Continuous	A1	A2	A3
Categorical	B1	B2	B3
Combination	C1	C2	C3

Category A1 includes two-level HLMs as well as multilevel factor and IRT models (Fox & Glas, 2001; Goldstein & Brown, 2002; Grilli & Rampichini, 2007). The previously discussed multilevel mixture IRT models proposed by Cho and Cohen (2010) and Vermunt (2008b), along with the two multilevel mixture factor models

proposed by Vermunt (2007) and Varriale and Vermunt (2012) are all from Category A2. These models assume a continuous latent trait at the person level, while introducing latent classes at the group level to cluster groups in terms of model parameters for the lower-level units. The idea of classifying groups is also applied to growth mixture models (Muthén, 2004), and its multilevel extension, MGMM-B, discussed by Palardy and Vermunt (2010), is a special case from Category A3. The multilevel mixture growth models classify both person and group units into homogeneous classes in terms of their mean growth trajectories. One type of multilevel latent class analysis (MLCA) from Category B2 introduces categorical latent variables at both the lower and higher levels (Asparouhov & Muthén, 2008; Vermunt, 2003). The higher level units are clustered based on the lower-level class membership probabilities. Vermunt (2003) and Van Horn et al. (2008) propose another type of MLCA from Category B1 with continuous random effects at the group level. These two approaches to specifying multilevel latent class are discussed in details in a later section of this dissertation.

The specification is complex for models in the C categories since those models introduce both continuous and categorical latent variables at the person level while considering latent variables at the group level. Allua (2007) proposed a multilevel variant of the factor mixture model of Category C1. The MMIRT models focusing on the possible procedure to identify school level DIF effect (Cho & Cohen, 2010; Vermunt, 2008b) are from Category C2 in which school units are clustered into group-level latent classes.

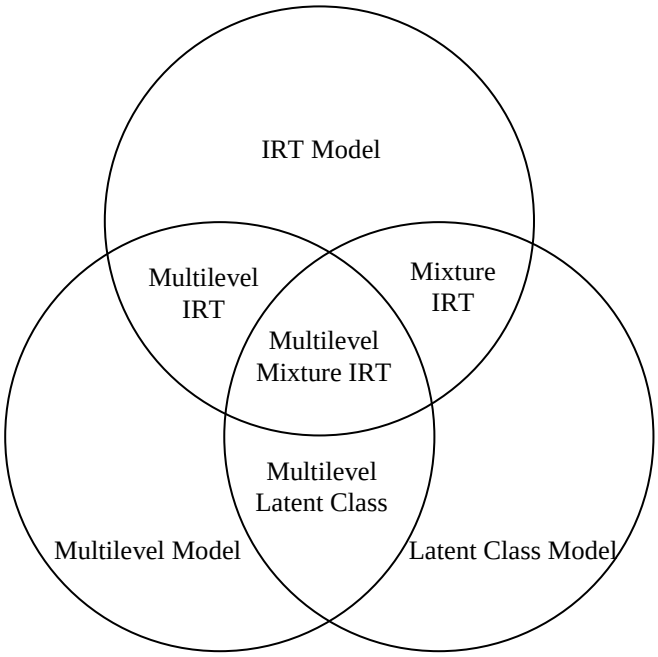

Figure 2.1. The conceptual relation between latent variable models (Cho, 2007)

As far as IRT models are concerned, many that belong to Category A2, A3, and C can be seen as special cases of the general MMIRT modeling framework proposed in the current study. As discussed in Cho (2007), the MMIRT integrates mixture IRT, multilevel IRT and multilevel latent class models. The Venn diagrams in Figure 2.1 depict the relations between these modeling approaches.

To date, the primary focus of MMIRT models introduced previously is to detect school-level DIF effect, and latent classes are introduced at the school-level. For instance, a discrete MMIRT model described by Cho and her colleague (Cho, 2007; Cho & Cohen, 2010) aims to identify school-level latent classes which present difference on item functioning. The authors claimed that the school-level DIF was a result of curricular or pedagogical differences (Cho & Cohen, 2010). While Cho's

model specified DIF effect on both student-level and school-level, Vermunt (2008b) proposed a variation of Cho's model in which only school-level DIF was considered.

Unlike the models proposed by Cho (2007) and Vermunt (2008b) which both focus on the possible procedure to identify school-level difference on item functioning, the current study emphasizes distinguishing between continuous and discrete distributions of variation at the group level in MMIRT. Two restrictive MMIRT models are introduced where the group-level random effects are modeled as either continuous or discrete. The two new models can be utilized to detect DIF when data are hierarchical. The method of distinguishing between the two modeling approaches may find support from the general discussion of the relation between the categorical and continuous latent variables.

In the following sections, a brief review of the three fundamental models is provided first, followed by the discussion of how MMIRT models are derived by combing these three models.

2.2 Mixture IRT Models

Mixture IRT models represent the integration of finite mixture models with conventional IRT models. Compared to conventional IRT models which use only continuous latent traits to represent the common content of observed responses, mixture models include a categorical latent variable to indicate the class membership of each examinee. These models assume that data arise from possibly heterogeneous populations consisting of several latent classes and a continuous latent trait can be incorporated to model the observed responses within each class. The discrete nature

of classes in finite mixture models facilitates interpretations of response differences in terms of latent class membership rather than manifest variables measured a priori.

Mixture IRT models provide sound solutions for detecting latent subpopulations that differ systematically on item responses. The early development of mixture IRT model started with the mixed Rasch model (Rost, 1990; 1997) that can identify items with different parameters across latent classes. Other variations such as the mixture linear logistic test model and mixture nominal model were utilized to identify examinees with random guessing behavior (Mislevy & Verhelst, 1990), or to detect differences in selecting response categories (Bolt, Cohen, & Wollack, 2001). Test speededness can be modeled using the mixture Rasch model with ordinal constraints (Bolt, Cohen, & Wollack, 2002).

The presence of DIF implies existing nuisance dimension(s) that cannot be captured by conventional latent variable models which assume a single latent trait. Therefore, Kelderman & Macready (1990) combined the ideas of latent class models and latent trait models, and suggested the use of loglinear latent class model to detect DIF by investigating interaction effect between grouping variables (either manifest or latent) and item parameters. Later development employed mixture IRT models to identify differential functioning of items (Cohen & Bolt, 2005; Cohen, Gregg & Deng, 2005; Samuelsen, 2005). Mixture IRT models can help researchers understand the causes of DIF by classifying examinees into latent classes. The new method also allows researchers to investigate the association between manifest variables and latent class membership. This is done by incorporating manifest variables as covariates.

The MMIRT models proposed in the current study are extensions of the mixture Rasch model (MRM). The assumption underlying the MRM is that a population consists of a fixed number of latent classes within which a Rasch model holds. Item difficulty parameters are allowed to vary across latent classes, but for members of one particular class all items function exactly the same. This mixture model not only quantifies latent ability but also accounts for qualitative differences among examinees. In the MRM, both item difficulty parameters and ability parameters get an extra subscript to indicate the latent classes they belong to.

2.3 Multilevel IRT Models

Traditional IRT models have been expanded in many ways to address methodological and empirical problems. One example is to specify an IRT model as a two-level model with items nested within examinees. Adams, Wilson, and Wu (1997) and Raudenbush and Sampson (1999) formulated a Rasch model within a hierarchical structure as a two- and three-level hierarchical logistic regression model. In this model, the first level specifies the relation between observed responses and latent ability. Within a hierarchical generalized linear model framework, Kamata (2001) proposed multilevel formulation of the Rasch model. Maier (2001) also described a Rasch model with a hierarchical model imposed on person parameters. Fox and Glas (2001, 2003) and Fox (2005) not only imposed a multilevel model on the two-parameter normal ogive model, they also included covariates at both levels as predictors of latent abilities. This type of reformulation is capable of modeling measurement error within and between item and examinee levels (Adams, Wilson, & Wu, 1997; Kamata, 1998). In addition, such modeling approach also provides more

accurate estimation of the standard errors of the parameters (Adams et al., 1997; Fox, 2005; Maier, 2001, 2002). More importantly, the combination of multilevel models with IRT leads to the increasing development of psychometric models for data with a hierarchical structure.

The multilevel IRT model has received more attention than the traditional multilevel models to investigate contextual effects (Fox, 2005). Rather than assuming a two-level structure, the multilevel IRT models impose a hierarchical linear model on the ability parameter. The models proposed by Kamata (2001) and Maier (2001) are both flexible to accommodate a third level (e.g., schools) and to further study its impact on the lower level (e.g., students) (Adams et al., 1997; Fox & Glas, 2001; Kamata, 2001; Maier, 2001, 2002). Cheong and Raudenbush (2000) specified a three-level multilevel IRT model to investigate school level impact on examinees' responses. The multilevel modeling framework can be utilized to detect DIF. The general procedure is to include covariates to account for the likelihood of a correct response that cannot be fully explained by latent ability (Wu, Adams, & Wilson, 1997). Cheong (2006) further extended the work of Wu et al. (1997) to a three-level model and investigated influences of school contexts on item performance differences across schools. DIF, thus, is interpreted as a significant cross-level interaction between item difficulty and individual and group characteristics (Cheong, 2006).

The Rasch hierarchical measurement model (HMM) proposed by Maier (2001) provides a foundation for modeling dichotomous responses within a nested structure. More specifically, the Rasch HMM incorporates a Rasch model and a two-

level hierarchical linear model and specifies intercepts as random effects at the first level. No additional covariates, however, are included at either level in this model.

2.4 Multilevel Latent Class Analysis

The latent class model is a statistical method for identifying unobservable groups of individuals (McLachlan & Peel, 2000; Muthén & Shedden, 1999). The main goal of using a latent class model is to construct meaningful clusters inferred from multiple observations. Traditionally, latent class models were developed for analyzing multivariate response data sets (Goodman, 1974; Lazarsfeld & Henry, 1968). Those models can be, however, conceptualized as a two-level model where the single-level multivariate responses are treated as two-level univariate responses with item responses nested within individuals (Vermunt, 2010).

As described by Vermunt (2008a) and Muthén and Asparouhov (2009), MLCA is akin to a mixed-effects regression model for categorical outcomes (Hedeker, 2003, 2008; Wong & Mason, 1985) which is latent rather than observed. Traditionally, a logistic regression model is used for a binary outcome. In MLCA, this outcome represents latent class membership. Conventional latent class models assume that the observations are independent of one another. This assumption, however, is often violated in many data such as when students are nested within schools or classrooms, or employees nested within companies. Thus, multilevel extensions of latent class models are proposed in response to the violation of independence assumption. If the traditional latent class analysis is conceptualized as a two-level model, a MLCA model has three levels where the nested structure is acknowledged by specifying intercepts of level-2 latent class as random effects. These random

intercepts allow the probability of membership in a particular level-2 latent class to vary across level-3 units and thereby to assess the influence of level-3 units on indicators that define level-2 latent class membership (Henry & Muthén, 2010).

Two approaches have been proposed to capture variation of latent class model parameters across group-level units. One variant of MLCA yields a clustering of higher-level units with regard to differences on lower-level responses or class membership probabilities. Another variant makes use of random effects as in conventional hierarchical linear models. Compared to the two-level latent class models, a MLCA includes either a discrete latent variable or continuous random effects at level 3 (Vermunt, 2010). The selection of discrete or continuous specification for the latent variables at level 3 depends on specific research purposes. However, Vermunt (2008a) advocated that the discrete approach shows more substantive benefit than the continuous approach. In the following sections, the situations where level-3 heterogeneity is modeled using continuous random effects or discrete latent variables are discussed first. The incorporation of covariates is also addressed in the later section.

2.4.1 Continuous Approach to MLCA.

The use of continuous random effects representing between-unit variation has been commonly adopted in a regression context (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). However, the inclusion of random effects in the estimation of mixture models remains understudied. Vermunt (2003, 2008a) and Asparouhov and Muthén (2008) have described mixture models with random effects in which the

groups are assumed to be drawn from a population of groups, and the probabilities of latent class membership are treated as random variables (Snijders & Bosker, 1999).

Taking multilevel structure into account, the most general multilevel latent class model assumes that all model parameters can be group-specific. The resultant model is equivalent to an unrestricted multiple-group latent class model (Clogg & Goodman, 1984). A more practical approach is to place restrictions on the general model by assuming that the item conditional probabilities are invariant across groups. This specification has been widely adopted in practice and is also employed here.

2.4.2 Discrete Approach to MLCA.

Rather than using continuous random effects, it is also possible to cluster higher-level units into one of several higher-level latent classes. Put differently, a second latent class model is imposed at the group level in addition to a person-level latent class model. This discrete approach to MLCA has been proposed by Vermunt (2003, 2008a) and Asparouhov and Muthén (2008).

Because the probabilities of person-level latent class membership are allowed to vary across groups, it is this variation that defines the between-group latent classes (Henry & Muthén, 2010). Instead of assuming a normal distribution of random effects, this assumption is replaced by a multinomial distribution (Vermunt, 2008a) with discrete latent values in the discrete approach. This is akin to using a discrete distribution in the form of a histogram to approximate a continuous distribution. Essentially, this approach relaxes the strong assumption pertaining to the form of the random effect distribution. This is advantageous to allow the presence of non-normality and to be less computationally demanding (Muthén & Asparouhov, 2008).

The discrete MLCA models use finite numbers of group-level latent classes to capture the group-level variability in the distribution of person-level latent class membership probabilities (Henry & Muthén, 2010). The identified higher-level latent classes are assumed to differ with respect to the probabilities of lower-level latent class membership. Consequently, a particular group-level latent class consists of groups with similar distribution of person-level typologies.

2.5 Multilevel Mixture IRT Models and Two Restrictive Cases

A general MMIRT model can be seen as a three-level model, where items are nested within examinees that are further nested within classrooms or schools. The level-1 model is concerned with item-level. The latent ability and latent class membership are modeled at level-2, the person-level. The level-3 model defines the variation of ability and probability of class membership across group units. MMIRT models enable researchers to investigate heterogeneity in individual response patterns while taking the multilevel data structure into account. Thus, individual responses on items are directly modeled as a function of not only individual characteristics but also the features of groups which the individuals belong to. In particular, other than using a set of latent classes combined with a continuous latent variable at the person-level, MMIRT allows probabilities of individual latent class to vary across higher-level units. That is, the probability that an individual will belong to a certain latent class is large in some groups while small in others. The specification of multilevel latent class models thus can be readily incorporated into MMIRT. More specifically, the random effects at the higher level are treated as either continuous or discrete in the same way as in MLCA.

Although two comparative approaches exist to modeling group-level random effects, the association between responses at the person-level is specified similar to a combination of mixture IRT and multilevel IRT. Mixture IRT models can capture heterogeneity of individual response patterns and help to infer the unobservable cause of DIF. The item difficulty portion of MMIRT together with the ability portion is built upon the conventional MRM (Rost, 1990). The conventional mixture IRT model is deficient in accounting for the nested structure as found in most educational data. Describing a latent trait in a multilevel IRT fashion is therefore adopted in the current MMIRT models. However, the decomposition of total ability variance into person-level and group-level components may not be practical in MMIRT. This is due to the fact that the distribution of ability is class-specific but the proportions of person-level latent classes are allowed to vary across group-level units.

In the following two sections, the integration of the MRM, multilevel IRT model and MLCA into the two approaches to MMIRT models is described first. In addition, covariates can be incorporated in the probability portion of the model to predict latent class membership at the person and group levels. How covariates from different levels are incorporated in MMIRT is illustrated in the third section. The exploration of covariate effects in MMIRT is beyond the scope of the current study. However, given the importance of covariates in the study of context effects, it is still worthwhile to briefly introduce the idea of modeling covariate effects in MMIRT. Two restrictive MMIRT models are proposed in the last section to answer one particular question, whether context effect affects the probability of individuals being clustered into a particular latent class.

2.5.1 Continuous Approach to MMIRT Model.

One substantial difference between continuous MMIRT and discrete MMIRT models is the specification of variation at the group level. The continuous MMIRT assumes that the groups are drawn from a population of groups. The model parameters are conditional on the particular group.

Let *g* denote a person-level latent class, $g = 1,...,G$, C_{jt} denote latent class membership for examinee *j* ($j = 1,...,J$) from group *t* ($t = 1,...,T$), and the probability that the examinee *j* belongs to the particular latent class *g* conditional on group *t* is denoted by $P(C_{jt} = g \mid T = t) = \pi_{g|t}$. Note that the group here refers to a class or school, rather than a manifest grouping variable such as gender and ethnicity. In a continuous MMIRT model, the unconditional probability of a correct response on item *i* ($i = 1,...,I$) is defined as

$$\begin{aligned} f(Y_{ijgt}) &= \sum_{g=1}^{G} P(C_{jt} = g \mid T = t) f(Y_{ijgt} \mid C_{jt} = g, T = t) \\ &= \sum_{g=1}^{G} \pi_{g|t} P(Y_{ijtg} \mid \theta_{jtg}, g, t, b_{ig}) \end{aligned} \quad , \qquad (2.1)$$

where Y_{ijgt} is the response to item *i* for examinee *j* from group *t* within latent class *g*, θ_{jtg} is the latent ability and b_{ig} is the item difficulty parameter for item *i* in latent class *g*. The conditional probability is written in the similar form as in a traditional Rasch model as

$$\begin{aligned} f(Y_{ijtg} \mid C_{jt} = g) &= P(Y_{ijg} = 1 \mid \theta_{jg}, g, t, b_{ig}) \\ &= \frac{\exp(\theta_{jtg} - b_{ig})}{1 + \exp(\theta_{jtg} - b_{ig})} \end{aligned} \quad . \qquad (2.2)$$

Item Difficulty Structure. The item difficulty parameters b_{ig} have no group subscript, indicating items function constantly across groups but differ across person-level latent classes.

Ability Structure. The latent ability θ_{jtg} is assumed to follow a normal distribution that is conditional on the person-level latent classes

$$\theta_{jtg} \sim N(\mu_g, \sigma_g^2), \tag{2.3}$$

where μ_g and σ_g^2 are the class-specific mean and variance, respectively. Given varying proportions of person-level latent classes in each group, to decompose the ability variation as specified in multilevel IRT models is not further carried out.

Probability Structure. The probability of class membership is defined as

$$P(C_{jt} = g \mid T = t) = \pi_{g|t} = \frac{\exp(\beta_{0tg})}{\sum_{p=1}^{G} \exp(\beta_{0tp})}. \tag{2.4}$$

with

$$\sum_{p=1}^{G} \pi_{p|t} = 1. \tag{2.5}$$

Since the latent class probability cannot be specified independently, knowing the probabilities of *G*-1classes automatically determines the probability of the last class. As a result, the model is nonidentifiable. For identifiability purpose, the first latent class is selected as reference group, and

$$\text{logit}\left(\frac{\pi_{g|t}}{\pi_{1|t}}\right) = \beta_{0tg}, \tag{2.6}$$

where β_{0tg} is the group-specific log odds of examinees belonging to latent class *g* instead of the first latent class conditional on group *t*, and for the first latent class $\beta_{01t} = \text{logit}(1) = 0$. The intercept parameter implies that the probability of individual

class membership is constant within each group. It is a random effect that captures the variability in the log-odds across groups.

The random intercepts can be divided into two components at the group level

$$\beta_{0tg} = \gamma_{00g} + U_{0tg}, \tag{2.7}$$

where γ_{00g} is the population average of the log odds for latent class *g* and U_{0tg} is the group-specific random deviation from the average of latent class *g*. Again, constraints such as $\gamma_{001} = U_{01t} = 0$ have been placed for identifiability. These random deviations are assumed to be normally distributed. The magnitude of the U_{0t} variance indicates the strength of the influence of the group level (Henry & Muthén, 2010). A larger variance indicates greater group effect.

For a total of *G* latent classes at the person level, *G*-1 random intercepts are specified with one class being selected as reference group. Each random intercept then requires a class-specific random variable to indicate the variability across groups. Unfortunately, this model becomes increasingly computational burden with growing number of level-2 latent classes (Van Horn et al., 2008; Vermunt & Van Dijk, 2001). Following the work of Bock (1972) and Hedeker (1999), Vermunt (2003) suggested modeling the means and covariances associated with the random variables using a common factor. Equation 2.7 can then be reformulated as

$$\beta_{0tg} = \gamma_{00g} + \tau_{00g} \cdot r_{00t}, \tag{2.8}$$

for $g = 2,...,G$, where τ_{00g} are factor loadings and r_{00t} is a normally distributed random effects with mean of 0 and variance of 1. For identifiability, $\gamma_{001} = \tau_{001} = 0$. The implicit assumption underlying this formulation is that the random means are

highly correlated and can be well represented by a single factor with different factor loadings for different random means (Asparouhov & Muthén, 2008; Vermunt, 2003). This factor model reduces the dimensionality of random means from (*G*-1) to 1 by specifying zero residual variance, and saves substantial amount of computation time. Van Horn et al. (2008) further suggested using a covariance structure with a (*G*-1)-dimensional multivariate normal distribution to relax this rather restrictive assumption.

Notice that this specification of MMIRT operates under the assumption of measurement equivalence, meaning that the model parameters for the response variables do not vary across groups (Vermunt, 2010). Groups differ only with respect to the probabilities of person-level latent class membership rather than their difference on item functioning.

2.5.2 Discrete Approach to MMIRT Model.

In discrete MMIRT models, rather than employing continuous random effects, mixtures are introduced at both the person level and the group level, each of which could capture a different type of unobserved heterogeneity (Vermunt, 2008a). Model parameters get one extra subscript to indicate the latent class that a group belongs to.

Following the subscripts used previously, let Y_{ijtgk} denote a specific item response. Notice that there are two types of identification, manifest (such as item *i*, examinee *j* and group *t*) and latent (such as person-level latent class *g* and group-level latent class *k*). Let *k* denote a particular group-level latent class, $k = 1,...,K$, C_t denote the latent class membership for group *t*, the probability that the group belongs to the particular latent class *k* is denoted by $P(C_t = k) = \pi_k$. Unlike the continuous

approach, the lower-level latent class membership of examinee *j* in group *t*, C_{jt}, is conditional on the higher-level latent class *k* rather than the group *t* as specified in the continuous approach, and the probability is then defined as $P(C_{jt} = g \mid C_t = k) = \pi_{g|k}$. The unconditional probability of a correct response on item *i* ($i = 1, ..., I$) is

$$\begin{aligned} f(Y_{ijtgk}) &= \sum_{k=1}^{K}\sum_{g=1}^{G} P(C_t = k)P(C_{jt} = g \mid C_t = k)\, f(Y_{ijtgk} \mid C_{jt} = g, C_t = k) \\ &= \sum_{k=1}^{K}\sum_{g=1}^{G} \pi_k \pi_{g|k} P(Y_{ijtgk} \mid \theta_{jtgk}, g, k, b_{igk}) \end{aligned} \quad , \qquad (2.9)$$

where the product of π_k and $\pi_{g|k}$ replaces $\pi_{g|t}$ as specified in the continuous model. The conditional probability is

$$\begin{aligned} f(Y_{ijtgk} \mid C_{jt} = g, C_t = k) &= P(Y_{ijg} = 1 \mid \theta_{jtgk}, g, k, b_{ik}) \\ &= \frac{\exp(\theta_{jtgk} - b_{igk})}{1 + \exp(\theta_{jtgk} - b_{igk})} \end{aligned} \quad . \qquad (2.10)$$

Item Difficulty Structure. The item difficulty parameter b_{igk} in Equation 2.10 is conditional on person-level latent class *g* and group-level latent class *k*. This is a more general specification.

Ability Structure. Similar with the mixture Rasch model, the latent ability level θ_{jtgk} is also assumed to follow a normal distribution,

$$\theta_{jtgk} \sim N(\mu_{gk}, \sigma^2_{gk}) , \qquad (2.11)$$

where μ_{gk} and σ^2_{gk} are the class-specific mean and variance, respectively. The subscripts for the means and variances indicate that they are allowed to differ across person-level latent classes conditional on the group-level latent classes. That is, for two group-level latent classes each of which has two person-level latent classes,

2 ×2=4 distinctive normal distributions can be obtained for latent ability. As discussed previously, the class-specific ability variance cannot be simply decomposed into individual- and group-level components, given varying proportion latent classes across groups.

Probability Structure. The conditional probability of latent class membership is specified in a way similar to continuous MMIRT with the only difference that the model parameter is conditional on the group-level latent class. The group-specific log odds of examinees belonging to latent class *g* instead of the first latent class conditional on group *t*, β_{0gk} is

$$\text{logit}\left(\frac{\pi_{g|k}}{\pi_{1|k}}\right)=\beta_{0gk}, \tag{2.12}$$

and $\beta_{01k}=0$. The intercepts are allowed to differ across latent classes of groups and is the random-effects portion of the model.

The probability of latent class membership at the group-level is specified as

$$P(C_t=k)=\pi_k=\frac{\exp(\gamma_{00k})}{\sum_{q=1}^{K}\exp(\gamma_{00q})} \tag{2.13}$$

with

$$\sum_{q=1}^{K}\pi_q=1, \tag{2.14}$$

and

$$\text{logit}\left(\frac{\pi_k}{\pi_1}\right)=\gamma_{00k}, \tag{2.15}$$

where γ_{00k} is the log odds of group *t* belonging to the higher-level latent class *k* instead of the first class, and for identifiability, $\gamma_{001}=\text{logit}(1)=0$.

2.5.3 Covariate Effect in MMIRT.

Mixture models benefit from incorporating covariates. First, covariates can help identify and describe characteristics of class membership. Several studies have shown that the use of covariates can improve detection of latent classes (e.g., Smit, Kelderman, & van der Flier, 1999; Cho, Cohen, & Kim, 2006). The use of covariates also helps to relieve the rigid requirement of latent class structure. In order to separate latent classes, mixture models require either substantial differences between latent groups or relatively large sample size. A simulation conducted by Smit et al. (1999) indicated that incorporating collateral information in MRM can substantially improve the estimation of standard errors and the assignments of latent classes. Recent studies employed covariates to formulate plausible explanations of the differences across latent classes on DIF items. For instance, Dai (2009) modeled a covariate effect directly in the mixing proportions in a mixture IRT model. The results indicated that the inclusion of covariates provided extra context information and achieved better recovery of the underlying structure.

In MMIRT models, the specification of covariates can be on both person-level and group-level. Covariate effects can differ across group units. As such, persons with same person-level covariate values can have different probabilities of being in a particular latent class due to contextual or environmental differences.

Person-level covariates are included to predict membership in person-level latent classes through multinomial logistic regression in both the continuous approach and discrete approach. Group-level covariates, in contrast, are specified differently between the two. For the continuous approach the group-level covariates are specified

using a linear regression function and are used to predict a group-specific probability that an individual belongs to a particular person-level latent class. The function of group-level covariates in the discrete approach can be either to predict the group-level latent class membership, or to predict person-level latent class membership. Both require the specification via a multinomial logistic regression.

Covariate Effects in Continuous Approach. Suppose a set of person-level covariates X_{rj} ($r = 1,...,R$), the class probability proportion of examinees, $\pi_{g|t}$ in the continuous approach is formulated as

$$\text{logit}\left(\frac{\pi_{g|t}}{\pi_{1|t}}\right) = \beta_{0tg} + \sum_{r=1}^{R} \beta_{rtg} X_{rjt}\,, \tag{2.16}$$

where β_{rtg} refers to the group-specific regression parameter and it can be treated as fixed effect as well as random effects across groups, and $\beta_{rt1} = 0$.

Given a set of group-level covariates W_{st} ($s = 1,...,S$), the group-level covariates are specified using a linear regression function as

$$\beta_{0tg} = \gamma_{00g} + \sum_{s=1}^{S} \gamma_{0sg} W_{st} + U_{0tg}\,, \tag{2.17}$$

where γ_{0sg} is the class-specific regression parameter for covariate W_{st} and $\gamma_{0s1} = 0$.

A graphic representation of continuous MMIRT is shown in Figure 2.2 modified from Henry and Muthén (2010, p.197). In this example, there are a total number of *G* person-level latent classes (C_g). The two black dots represent the random means for the person-level latent classes. As explained above, there are *G*-1 random means (therefore, *G*-1 filled circles) for *G* person-level latent classes. In addition, these random means are allowed to correlate with one another.

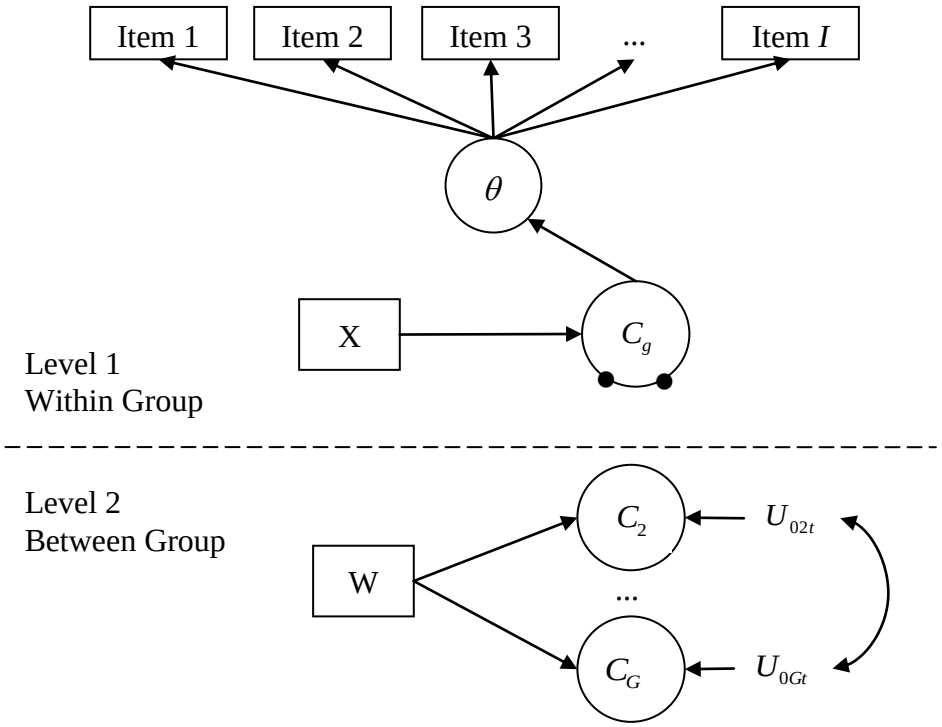

Figure 2.2. Multilevel mixture IRT model -- continuous approach

Covariate Effect in Discrete MMIRT. Similar to the continuous approach, the person-level covariates are included to predict the probability $\pi_{g|k}$ in the discrete approach. The group-level covariates can directly predict the person-level latent class membership. In addition, another set of group-level covariates indirectly impact person-level class by directly predicting the group-level latent class membership.

Suppose *R* person-level covariates X_{jt} and *L* group-level covariates W_t, the equation for $\pi_{g|k}$ is written as

$$\text{logit}\left(\frac{\pi_{g|k}}{\pi_{1|k}}\right) = \beta_{0gk} + \sum_{r=1}^{R}\beta_{rgk}X_{rjt} + \sum_{l=1}^{L}\beta_{0lg}W_{lt}\,, \qquad (2.18)$$

where β_{rgk} refers to the class-specific regression parameter for person-level covariates and can vary between the group-level latent classes, β_{0lg} is the class-specific

parameter for group-level covariates and is considered fixed. Again, additional constraints are placed, $\beta_{r1k} = \beta_{0l1} = 0$.

Covariate effects at the group-level are specified via a multinomial logistic regression. For another set of *S* group-level covariates $W_t^{'}$

$$\text{logit}\left(\frac{\pi_k}{\pi_1}\right) = \gamma_{00k} + \sum_{s=1}^{S} \gamma_{0sk} W_{st}^{'} , \qquad (2.19)$$

where γ_{0sk} is the class-specific regression parameter of a covariate $W_t^{'}$ for the group-level latent class *k*, with the constraint $\gamma_{0s1} = 0$.

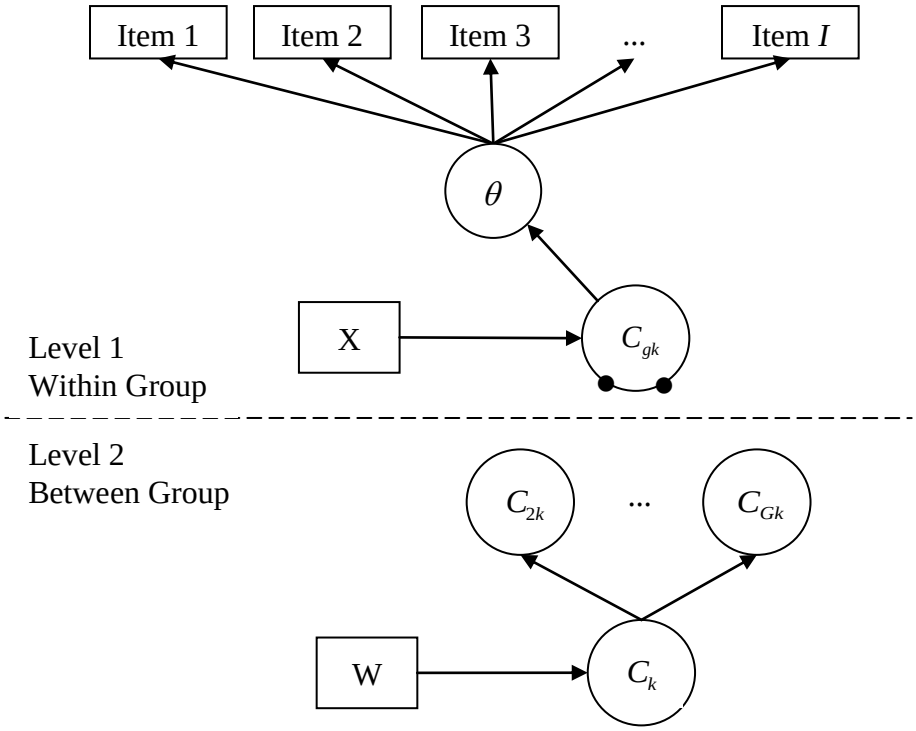

Figure 2.3. Multilevel mixture IRT model -- discrete approach

A graphic representation of discrete MMIRT is shown in Figure 2.3. Assume that *G* person-level latent classes (C_{gk}) and *K* group-level latent classes (C_k) exist in the sample. Again, two black dots are used to represent the *G*-1 random means for the

person-level latent classes. Those random means are conditional on the k^{th} group-level latent class. In addition, the effect of group-level covariates on the person-level latent classes is not presented in the graph below.

2.5.4 Two Restrictive MMIRT models.

More often, restrictions are placed on the general model for specific research purposes. In the current study, modeling the variation in probabilities of person-level class membership is the main focus in the models proposed below. In this section, two restrictive MMIRT models, which differ only at the specification of group-level variation, are further described.

MMIRT models proposed by Cho and her colleague (Cho, 2007; Cho & Cohen, 2010) used discrete latent classes at the group level to capture between-group differences on item difficulties. A more restrictive model is obtained by assuming that the item parameters do not depend on the group-level unit. Following the notation used before, this means $b_{igk} = b_{igk'}$ for $k \neq k'$. This notation indicates that the item difficulty parameters differ only across the person-level latent classes. Latent classes at the person level capture the heterogeneity in response patterns, whereas latent classes at the group level differ in terms of the probability of individuals being classified in a particular person-level latent class. Put differently, the group-level latent classes have different distributions of random probabilities of person-level classification. An assumption of multinomial distribution replaces the normal distribution assumed in the continuous approach (Vermunt, 2008). Group-level latent classes represent a discrete distribution in the form of a histogram, where

nonnormality is allowed (Henry & Muthén, 2010). Thus, two MMIRT models differ merely on whether the group-level variation is specified as continuous or discrete.

In addition, although latent ability is allowed to follow distinctive distributions within latent classes, the current restrictive models define the ability distribution in the form of Rasch HMM. Such model constraints provide a practical benefit for the current study for it allows further to decompose the variation of latent ability into between-group and within-group components. More specifically, a two-level hierarchical linear model is further imposed to model variation of the latent ability within and between group units. The examinee's ability is specified as the sum of a fixed effect and a random effect

$$\theta_{jt} = \mu_{0t} + u_{jt} \tag{2.20}$$

where μ_{0t} is the mean ability of group *t*, and u_{jt} is the ability variation within groups.

Rasch HMM is a special case of random-effects logit models, the individual random effects u_{jt} are assumed to follow a logistic rather than a normal distribution as commonly seen in linear multilevel models (Rodríguez & Elo, 2003). To be precise, the logistic regression is assumed to have mean 0 and variance $\sigma_{0t}^2 = \frac{\pi^2}{3} s^2$ where *s* is the location parameter. The group-level model for ability is specified as

$$\mu_{0t} = \mu_{00} + \nu_{0t}\,, \tag{2.21}$$

where μ_{00} is the grand mean ability, and ν_{0t} is the between-group ability variation and follows a normal distribution with mean of 0 and variance of σ_{00}^2 (i.e., $\nu_{0t} \sim N(0, \sigma_{00}^2)$). Similar to conventional linear multilevel models, the intra-class correlation (ICC) is

utilized to indicate the proportion of variance explained by group units (Rodríguez & Elo, 2003). Thus, the ICC in Rasch HMM is

$$ICC = \frac{\sigma_{00}^2}{\sigma_{00}^2 + s^2\pi^2/3}. \quad (2.22)$$

In brief, both MMIRT approaches are capable of detecting and describing characteristics of group heterogeneity, while accommodating the hierarchical data structure. In addition to explore potential DIF, the MMIRT methods facilitate simultaneous description of mixtures at the group level. The continuous approach captures the variation between groups using normally distributed random effects. In contrast, the discrete approach seems to offer substantive benefits as it does not require making as strong assumptions about the distributions of random effects as does the continuous approach and is less computational demanding (Muthén & Asparouhov, 2008; Vermunt, 2008a). If substantial difference is assumed among groups, to identify group-level latent classes by specifying relevant model parameters to be class dependent is a proper solution. However, such specification requires a strong theoretical rationale to support.

After imposing certain constraints, the two restrictive MMIRT models proposed above differ only in terms of the distributions of group-level variation. In that sense, a model comparison perspective can be adopted to decide which of the two approaches is better at describing the underlying distribution. Given the absence of evidence in existing literature, model comparison between two MMIRT approaches is based on the discussion on how to distinguish between categorical and continuous latent variables.

2.6 Distinguishing between Categorical and Continuous Latent Variables

A number of researchers (e.g., Bauer & Curran, 2004; Haertel, 1990; Heinen, 1996; Molenaar & von Eye, 1994; Reise & Gomel, 1995; Vermunt, 2001) have discussed extensively the relation between categorical and continuous latent variable models. Existing latent variable modeling framework provides a compelling approach to distinguish between nominal (i.e., class, qualitative) latent variables and metrical (i.e., real valued or interval) variables (Markon & Krueger, 2006). Nominal latent variable models are equivalent to the metrical latent variable model because nominal latent variable models can be accommodated by metrical latent variable models (Haertel, 1990; Molenaar & von Eye, 1994). This is similar to the use of dummy coding to accommodate analysis of variance models, which are nominal, and in regression models, which are metrical. Nominal latent variable models are not simple discrete metrical latent variable models in that they capture multidimensionality in latent structure. More precisely, nominal latent variable models are multidimensional discrete metric latent variable models and these two models fit the same datasets equally well (Haertel, 1990; Molenaar & von Eye, 1994).

Models representing either continuous or discrete distributions are not directly compared to each other to infer the discreteness versus continuousness of the data (Markon & Krueger, 2006). Among the metrical latent variable models, it is generally recognized that a continuous distribution can exactly reproduce discrete latent variable models (Haertel, 1990). A continuous latent distribution can conceptually reproduce discrete latent distribution because the possible latent values contained by a discrete latent distribution are subsumed by the latent values contained in a

continuous latent distribution (Markon & Krueger, 2006). The restrictive distribution assumption underlying the continuous approach, however, prevents its application in more general scenarios where non-normality may occur. In contrast, a discrete latent distribution is more flexible in its' distributional assumptions and is capable of approximating a continuous distribution with arbitrary precision (Heinen, 1996; Vermunt, 2001). For example, researchers (e.g., Aitkin, 1997; Vermunt & Van Dijk, 2001) have indicated that a finite mixture distribution can be obtained from the discretization of a continuous latent variable distribution. The approximation of continuous distributions gets better with increasing numbers of discrete values, suggesting a less fundamental distinction between continuous and discrete latent variables (Vermunt & Magidson, 2005). As Molenaar and von Eye (1994) remarked, the choice of continuous versus discrete scaling for the latent variables is essentially arbitrary as long as the analysis is confined to model means and covariances. The central question to ask is whether a limited number of latent values or a large number of latent values is required to account for an observed distribution.

In practice, the number of discrete latent values is relatively small, therefore model fit and parameter estimates can differ appreciably between a discrete variable model and its continuous counterpart (Haertel, 1990). This difference in fit essentially provides important information for decision making about which latent structure, continuous or discrete, should be selected for a particular set of data (Markon & Krueger, 2006). Lubke and Neale (2006) advocated the use of model fit to decide an underlying latent variable is continuous or categorical. Although the overextraction problem may occur when either fitting a continuous latent variable model to data

stemming from a heterogeneous population or fitting latent class models to data with continuous factors. Comparing the fit of different exploratory models usually leads one to correctly select between categorical and/or continuous latent variables. Class separation is found to have profound impact on model fit indices. In addition, within-class parameterization is better recovered with increasing within-class sample size, which leads to correct model selection (Lubke & Neale, 2006).

Model misspecification is one of the most important issues that arise in distinguishing between discrete and continuous latent structure. Most often, this problem is discussed with respect to normal and non-normal distributions, but it can certainly apply to a more general scenario where a continuous latent distribution model is severely misspecified (Markon & Krueger, 2006). A variety of methods have been proposed to resolve this issue (see e.g., Bauer & Curran, 2004; Maraun, Slaney, & Goddyn, 2003; Miller, 1996). Continuous latent variables are commonly assumed to follow a normal distribution. When a set of data is non-normally distributed, a discrete latent variable can capture the non-normality. Bauer and Curran (2004) illustrated the effect of non-normality on latent class estimation using simulated data. Their findings show that the presence of two latent classes better approximate a non-normal multivariate distribution, even when only one group truly exists in the population. In other words, for a sample from a non-normal population, model comparison may favor a discrete model with multiple values over a normal continuous model. The use of model fit to infer the correct number of classes may be misleading as the additional populations capture features of the non-normality (Bauer & Curran, 2004; Markon & Krueger, 2006).

Given the fundamental relation between discrete and continuous latent variables, the only advantage of the discrete specification is that this approach does not introduce possible inappropriate and unverifiable assumptions about the distributions of latent variables (Bauer & Curran, 2004; Vermunt, 2008a). However, it is also not necessarily true that a discrete model generally gives better approximations than a continuous non-normal distribution or normal distribution. Under certain conditions, a normal distribution itself might be preferred over a discrete latent distribution for a latent non-normal distribution because the normal distribution is associated with loss of less statistical information about the observed sample (Markon & Krueger, 2006). Moreover, the efficiency of approximation to the population model varies among different discrete latent variable models. Information-based fit criteria assess the amount of information lost in approximating an observed distribution by a model-generated distribution. Those fit indices would suit the purpose of distinguish between continuous and discrete latent variables.

To evaluate the performance of model fit indices in distinguishing continuous and discrete MMIRT models, a simulation study is conducted and detailed description is given in Chapter 3. Before the simulation design is described, technical issues with respect to estimation methods and information-based model fit indices are discussed.

Chapter 3: Methods

The first chapter described the motivation for adopting a multilevel mixture modeling framework and the connection between the two approaches. The second chapter provided the theoretical background and mathematic expression of two restrictive MMIRT models. This chapter addresses two issues of model estimation for MMIRT first, and later a simulation study is introduced to investigate the model selection between the continuous and discrete MMIRT models.

3.1 Estimation and Model Selection

3.1.1 Maximum Likelihood Estimation.

The unknown parameters of the MMIRT models described previously can be estimated by means of maximum likelihood (ML). ML estimates are consistent and can be developed for various estimation situations. ML methods also offer desirable mathematical and optimality properties, such as estimators are asymptotically unbiased with minimum variance as sample size increases and they approximate normal distributions and sample variances for hypothesis testing of the parameters. ML has been widely utilized to estimate the parameters that define statistical models, and in fact, is the gold standard to which other estimation methods are often compared.

The implementation of ML estimation in multilevel factor mixture models has been demonstrated by Vermunt and his colleagues (Varriale & Vermunt, 2012; Vermunt, 2003). The likelihood function described below extends Vermunt's equations to accommodate mixture IRT at the lower-level.

ML estimation is a process of finding the estimates for latent variables to maximize the likelihood function for observed responses. In a three-level model, suppose $\mathbf{Y}_t$ is the vector of observed responses of group *t* and $\boldsymbol{\eta}$ is the complete set of unknown parameters which are treated as fixed, the likelihood of the observed data marginal to all latent variables is

$$L = \prod_{t=1}^{T} f^{(3)}(\mathbf{Y}_t), \tag{3.1}$$

where $f^{(3)}(\mathbf{Y}_t)$ is the probability density of the observations of group *t*. The groups are assumed to be independent, and the product is over all group units.

For continuous latent variables at the group level, $f^{(3)}(\mathbf{Y}_t)$ is given by

$$f^{(3)}(\mathbf{Y}_t) = \int_{\boldsymbol{\eta}_t^{(3)}} \left[\prod_{j=1}^{J} f(\mathbf{Y}_{jt} \mid \boldsymbol{\eta}_t^{(3)}) \right] f(\boldsymbol{\eta}_t^{(3)}) d\boldsymbol{\eta}_t^{(3)}, \tag{3.2}$$

where $\boldsymbol{\eta}_t^{(3)}$ is the continuous latent variables at group level and $f(\mathbf{Y}_{jt} \mid \boldsymbol{\eta}_t^{(3)})$ is the conditional density of each person. The persons within group *t* are assumed to be independent given the random variables, $\boldsymbol{\eta}_t^{(3)}$.

When latent variables are discrete, the integration over $\boldsymbol{\eta}_t^{(3)}$ in Equation 3.2 is replaced by a summation over *K* group-level latent classes. The likelihood for group *t* is then defined by

$$f^{(3)}(\mathbf{Y}_t) = \sum_{k=1}^{K} \pi_k \left[\prod_{j=1}^{J} f(\mathbf{Y}_{jt} \mid \boldsymbol{\eta}_{kt}^{(3)} = 1) \right]. \tag{3.3}$$

where π_k is the class weight, and $f(\mathbf{Y}_{jt} \mid \boldsymbol{\eta}_{kt}^{(3)} = 1)$ is the density of one person conditional on the *k*th group-level latent class.

Under the current specification, $f(\mathbf{Y}_{jt} \mid \boldsymbol{\eta}_t^{(3)})$ and $f(\mathbf{Y}_{jt} \mid \boldsymbol{\eta}_{kt}^{(3)} = 1)$ are expressed by the similar form as in MRM. A detailed review of ML in MRM is presented by Formann (2007).

In solving the integrals involved in the computation of likelihood function, a closed form expression is available when responses and latent variables are normally distributed (Vermunt & Magidson, 2005). In other cases, numerical integration approximates an integral by a weighted sum of the integrand function. This function is evaluated by a set of quadrature points of the variable being integrated out. Skrondal and Rabe-Hesketh (2004) provided a comprehensive discussion about alternative approaches such as Laplace approximation and Monte Carlo integration to approximate the integrals.

To maximize the likelihood function, the Expectation-Maximization (EM) algorithm, Newton-Raphson (NR) algorithms and Fisher scoring algorithms are commonly implemented. The EM approach includes two steps: the E-step to evaluate the posterior expectation function, and the M-step to maximize this expectation function and update estimates of parameters. For ML estimation of discrete multilevel models with more than two levels, a new algorithm, which makes use of the conditional independence assumption, updates the expectation function upward and downward through the hierarchical structure (Vermunt, 2003). Compared to the EM algorithm, NR and Fisher scoring algorithms can produce estimates of standard errors for the maximum likelihood estimate. Both methods work in a very similar way, using the first-order derivatives and the second-order derivatives of the log-likelihood function. One difference is that NR uses the Hessian matrix in the place of Fisher's

information matrix. To overcome the problem encountered in computing the derivatives either analytically or numerically, other algorithms such as Quasi-Newton (QN) algorithms have been proposed.

Most multilevel mixture models introduced previously can be fitted using either of the two popular software packages - M*plus* (Muthén & Muthén, 1998-2010) and *Latent GOLD* (Vermunt & Magidson, 2005, 2008), which implement slightly different combinations of the integration and maximization methods. *Latent GOLD* solves the integrals using Gauss-Hermite integration, and uses the EM algorithm coupled with NR algorithm to find the ML estimates (Vermunt, 2010). Specifically, the estimation process starts with the upward-downward algorithm, and the NR algorithm takes over when the estimates approach the final solution (Vermunt & Magidson, 2005). M*plus* employs a similar procedure but using rectangular, Gauss-Hermite, or Monte Carlo integration for numerical integration and the optimization is achieved using a combination of EM and QN method (Muthén & Muthén, 2006). In particular, the M*plus* software allows using multiple random starting values to avoid local maximum problems and only the starting values with the highest log-likelihood among these runs are used as the starting values in the final stage of optimization. Both packages have options for obtaining robust standard errors as well as for dealing with missing values and complex sampling designs. The current study used M*plus* for MMIRT model estimation.

3.1.2 Information-based Model Fit Statistics.

Given the non-nested relation between a model of continuous latent distribution and a model of discrete latent distribution, the standard likelihood ratio

test is not suitable to assess relative model fit (McLachlan & Peel, 2000). Instead, the distribution-free information criterion statistics, which are based on the log-likelihood, are commonly used to compare these two types of latent distributions and to make inferences about population structure.

Although numerous information criteria exist, many can be seen as special cases of minimum complexity criteria (Barron & Cover, 1991; Sclove, 1987) that adjusts the log-likelihood for model complexity. Minimum complexity criteria have the general form written as

$$IC = -2\log[L(M)] + C(M), \tag{3.4}$$

where IC the value of a certain information criterion is a combination of $-2\log[L(M)]$, -2 times the log-likelihood of the model *M*, and $C(M)$, a quantity presenting the complexity of model *M* (Barron & Cover, 1991). The quantity, $C(M)$, reflects the amount of information required to describe model *M* and can be further presented as a product of $a(n) \times p$ (Sclove, 1987), where n is sample size, and p is the number of estimated parameters. In general, more parsimonious models, which are usually preferable, produce smaller values of minimum complexity criteria, $a(n)$ therefore is a penalty term added to the -2 log-likelihood for each additional estimated model parameter (Henson, Reise & Kim, 2007). However, models with more parameters are always found to fit the data at least well or even better, meaning a greater log-likelihood. As such the impact of the penalty could be cancelled out, resulting in the more complex model being favored.

Some important examples of minimum complexity criteria include the Akaike information criterion (AIC; Akaike, 1974, 1987), the Bayesian information criterion

(BIC; Schwarz, 1978), the sample size adjusted BIC (ssBIC; Sclove, 1987), the consistent AIC (CAIC; Bozdogan, 1993). These information criteria differ in the way they specify model complexity in terms of sample size and the number of free parameters of the fitted model. More specifically, the AIC does not depend on sample size and the penalty is $a(n) = 2$. The BIC, CAIC, and ssBIC criteria integrate sample size in different ways. Each additional parameter is penalized identically in the BIC and CAIC as for the BIC the penalty term is $a(n) = \log(n)$ and for the CAIC is $a(n) = \log(n) + 1$. Unlike in BIC and CAIC, the ssBIC penalizes complexity based on the Rissanen Information Criteria (Rissanen, 1978) for autoregressions, and the penalty term is $a(n) = \log\left(\frac{n+2}{24}\right)$. Bozdogan (1993) suggested a modified AIC (AIC3) criterion using 3 instead of 2 as penalizing factor to avoid negatively biased estimate of the expected Kullback-Leibler information in the fitted model (Hurvich & Tsai, 1989) as existing in the standard AIC. The same reasoning applies, another modification of AIC, the AICC proposed by Burnham and Anderson (2002) takes the ratio of sample size to model parameters into consideration.

$$AIC = -2\log[L(M)] + 2p \tag{3.5}$$

$$AIC3 = -2\log[L(M)] + 3p \tag{3.6}$$

$$AICC = -2\log[L(M)] + \frac{2np}{n - p - 1} \tag{3.7}$$

$$CAIC = -2\log[L(M)] + (\log(n) + 1)p \tag{3.8}$$

$$BIC = -2\log[L(M)] + \log(n)p \tag{3.9}$$

$$ssBIC = -2\log[L(M)] + \log\left(\frac{n+2}{24}\right)p \tag{3.10}$$

The BIC has been recommended for its consistency across a variety of modeling settings. This index tends to select the correct model more frequently as sample size increases (Haughton, 1988; Leroux, 1992). The BIC is more conservative than the AIC for selecting models with more parameters, and the CAIC is the most conservative. The difference in the CAIC compared to the BIC results in a preference for smaller models slightly more often than does the BIC. However, with sufficiently large sample size, the BIC and CAIC never lead to diverging results (Markon & Krueger, 2006).The penalty of additional parameters in the ssBIC is not as harsh as in the BIC, and only when sample size exceeds 176 will the ssBIC become larger than AIC (Henson, Reise & Kim, 2007). The ssBIC is advocated for better performance when the model has either a large number of parameters or a small sample size (Yang, 2006).

To distinguish continuous and the discrete latent variable models, AIC, BIC, CAIC and ssBIC are the four criteria most often used (e.g., Bauer & Curran, 2004; Lubke & Neale, 2006; Markon & Krueger, 2006). Lubke and Neale (2006), for example, found that when models with categorical, continuous, or both types of latent variables are fitted to the data generated under different types of latent variable models, correct model selection is more often made by the AIC and ssBIC. These two criteria outperform the BIC and CAIC.

In the context of mixture modeling, no conclusive results have been reached regarding the function of various information criteria. McLachlan and Peel (2000) noted that the AIC tends to overestimate the number of classes present, whereas the BIC (and by extension the CAIC) may underestimate the number of classes present,

particularly in small samples. Compared to the wide use of AIC and BIC, only one study (Dias, 2006) supported the use of AIC3 in finite mixture models for selecting the number of latent classes. The model comparison between continuous and discrete MLCA models suggested that the BIC might not function properly in multilevel mixture models (Henry & Muthén, 2010).

The information criteria used in the standard mixture analysis can also be utilized as model selection measures in multilevel mixture models. However, model selection becomes an even more complex issue for this type of model, especially when group-level heterogeneity is modeled using group-level latent classes (Vermunt, 2010), because the decision on the required number of latent classes not only has to be made at the person level, but also at the group level.

The use of criteria that contain sample size in their formula is particularly problematic, because the sample size can be measured in various ways. The sample size can refer to the number of observations both within- and between-levels. Palardy and Vermunt (2010) suggested using group-level instead of person-level sample size in BIC when comparing models that differ only at the group-level. A recent simulation study by Lukociene and Vermunt (2010) supported the use of the modified version of BIC. To evaluate the impact of change of sample size, the current study includes the modified BIC, as well as the other three fit indices with sample size information. The total person-level sample size is replaced by the number of groups for those indices. To differentiate them from the original indices with total person-level sample size, the letter "n" is added before the abbreviations of the modified indices, for instance nAICC, nCAIC, nBIC and nssBIC. The four modified fit indices,

combined with the six fit indices mentioned above are included in the current study to select best fitting models.

3.2 Simulation Design

The findings from previous studies on mixture IRT, MLCA, as well as comparison between continuous and discrete latent variable models can be utilized as foundations for the investigation of MMIRT models. MMIRT models can be seen as multilevel extensions of finite mixture models. As in most mixture models, the primary goal is to assign individuals to their most likely classes. The quality of class assignment at the person level plays an even more important role in MMIRT as it determines whether the group-level random effects can be identified successfully.

It is well established that the characteristics of measurement can impact the class assignment in mixture models. Effects of class separation that are due to the property of measurement instrument such as test length, magnitude of DIF effect, proportion of DIF items are frequently investigated in mixture IRT models. In general, clearer class separation can be achieved using a longer test containing a larger proportion of items with greater size of DIF effect. With respect to person features, one factor often assessed is the difference of latent ability distributions between classes (e.g., Cho & Cohen, 2010; Dai, 2009). Other than that, however, few studies in the literature have addressed the factors that are relevant to persons, groups and the interaction between these two levels of units, especially sample size at the two levels, ability variation within and between groups. Therefore, in the current study the characteristics of test and distribution of ability are held constant, while the other factors related to personal-level features were manipulated.

3.2.1 Fixed Factors.

A simulated measurement scenario was constructed with a test of 40 items. The number of items reflected a length commonly seen in educational tests. Item difficulty parameters in IRT were generated from a uniform distribution of $U(-1.5,1.5)$. To introduce DIF effect, a selected proportion of items were associated with difference on item difficulties between person-level latent classes.

In the context of DIF, two types of qualitative differences are identified (De Boeck, Wilson, & Acton, 2005). Simple qualitative differences refer to the condition where the location of item difficulties has a discernible pattern among the latent classes. In contrast, there is no such apparent pattern in the location of item difficulties in the case of complex qualitative differences. In the current study, the magnitude of DIF effect was fixed at 1 to reflect a simple qualitative difference.

In addition, a data set with a sample of 6000 individuals with a total variance of latent ability of 1 was simulated. The sample size reflects a grade size typically seen in a county. Two latent classes were assumed to exist at the person level. A summary of fixed factors was provided in Table 3.1.

Table 3.1 A summary of fixed factors

Factors	Model Level	Fixed Values
Test length	Item-level	40
Item difficulty range	Item-level	$U(-1.5,1.5)$
Total sample size	Person-level	6,000
Person-level mixtures	Person-level	Two
Total ability variance	Group- and Person-level	1

3.2.2 Manipulated Factors.

The primary goal of this simulation study is to assess the performance of information-based fit criteria in distinguishing between the continuous and discrete MMIRT models and establish the conditions under which practitioners can properly apply the two approaches. Both continuous and discrete distributions were used to generate random effects at the group-level. Special interest was to what extent the four factors: person-level class separation, within-group sample size, proportion of mixtures as well as group-level ability variance can affect the model identification. A summary of manipulated factors was provided in Table 3.2.

Table 3.2 A summary of manipulated factors

Factors	Model Level	Corresponding Values	
Percentage of DIF items	Item-level	15%	30%
Group sizes	Person-level	25	150
Proportion of mixtures	Person-level	50% : 50%	30% : 70%
Group-level ability variance	Group-level	0.1	0.3
Discrete Distribution			
Number of discrete values	Group-level	Two	Four
Continuous Distribution			
Distribution forms	Group-level	Normal	

Class separation. Whether it is easy or difficult to classify individuals to latent classes, this is a question concerning class separation. The current study investigated the effect of class separation due to various percentages of DIF items within a test. More precisely, a small proportion of items, such as 15%, is specified to function differentially across the two latent classes and is expected to result in weak class separation. In contrast, when 30% of items are assumed to have DIF effect, this

condition is considered to reflect large class separation. These two percentages are typically observed in educational assessments (Hambleton & Rogers, 1989; Raju, Bode, & Larsen, 1989). It was expected that a larger percentage of DIF items would lead to a better separation of classes at the person level.

Note that class separation interacts with sample size. For more dissimilar classes, smaller samples are required for class separation. In MMIRT, the separation between person-level latent classes may interact with the sample size both at the person- and group-level. It is unclear how the group-level sample size may affect the separation of person-level classes, if at all.

Person-level and group-level sample size. It was of interest to investigate MMIRT behavior with respect to the sample size requirement at person-level and group-level. Proper application of multilevel mixture models requires sufficient sample sizes at each of the three levels. Vermunt (2010) provided some general guidelines on the sample size requirement in MLCA. A simple cut-off value is impractical in these types of models because the sample size at one particular level affects not only sampling fluctuation but also the separation of latent classes at higher levels. When group-level latent classes are introduced in the model, the required person-level and group-level sample sizes depend heavily on the separation of the person-level and group-level classes, respectively.

To mimic the scenario regularly seen in statewide tests, two group sizes, 25 and 150 are selected to reflect the number commonly seen for a classroom and a grade within school. Given the fixed total sample size, the selection of group sizes

also determined the number of groups to be 240 and 40 accordingly, which reflect a large and small number of groups.

Proportion of person-level mixtures. Another factor that is often addressed in mixture modeling is the proportion of latent classes in the population. In the current study, this factor was only manipulated at the person level. Other relevant studies often used equally split latent classes (Smit et al., 1999; Cho, et al., 2007). It should be noted that since the person-level latent classes are indicators of group-level latent classes, varying proportion of mixtures at the person level could potentially impact the identification of group-level random effects. Therefore, other than the condition with 50%:50%, an additional 30%:70% proportion was included to reflect the condition with uneven proportion of classes. It was expected that the uneven proportion condition would lead to smaller variation at the group level, which further increased the difficulty to detect group heterogeneity.

Between-group ability variance. The simulation studies of conventional mixture IRT models have revealed that the recovery of class membership improves with more distinct latent groups. This is usually done by assuming that the mixtures are sampled from distributions with different population means. Instead, how the within- and between-group variation of latent ability can interact with the identification of random mixtures is of interest in the current study.

A between-group variance of 0.3 is selected to reflect that the group level variance accounts for 30% of total variance as commonly found in multilevel studies (Cheong, 2006; Palardy & Rumberger, 2008). As random-effects logit models, the

within-group variation of latent ability in MMIRT models is assumed to be $\frac{1}{3}\pi^2 s^2 = 0.7$, and the scale parameter is approximately 0.21.

Another variance of 0.1 is selected to reflect a more homogeneous condition where the group-level accounts for only 10% of total variation. Accordingly, the scale parameter for the within-group variation is around 0.27. The consequence of groups with homogeneous ability distributions is to force the identification of person-level latent classes to rely more on the difference in item functioning rather than on ability distribution within groups.

Random effect distributions. Both discrete and continuous MMIRT models are used to generate data. The ability of information-based model fit indices to identify population distribution of MMIRT models under different distributions of random effects is examined.

Equation 2.1 is used to generate the continuous distributions. When the population distribution is continuous, data sets are sampled from a normal distribution with mean 0 and variance 0.5, coupled with various sample sizes and levels of class separation. Each simulated data set is analyzed using four different models: the data-generation model and three discrete models with two, three and four discrete latent values.

Equation 2.9 is used to generate the discrete distributions. Distributions associated with few discrete values should be easily distinguished from continuous distributions. With increasing number of latent values, a discrete distribution and a continuous distribution become more indistinguishable (Markon & Krueger, 2006). Populations with two or four discrete values are employed for this condition. The

marginal probabilities of each group-level latent class are assumed to be equal, and the detailed specifications of mixture proportions are specified in Table 3.3. The selection of relative proportions within each cell must meet two criteria: first, to maintain the marginal probabilities of person-level and group-level latent classes, simultaneously; second, to introduce a moderate level of variation over conditional probabilities. Note that the relative frequencies within each cell indicate that the person-level latent classes and group-level latent classes are dependent. This is a result of identification of the group-level latent classes in discrete MMIRT models. The group-level latent classes describe the probability of membership in each person-level latent class (Henry & Muthén, 2010).

Four models are fitted to the same generated dataset: three discrete models including the data-generation model and a continuous model that assumed the random effect distribution to be standard normal (i.e., $N(0,1)$). Specifically, for the two-class condition the three discrete models have two to four group-level latent classes, and for the four-class condition the numbers are 3, 4 and 5.

Table 3.3 True probabilities of latent classes at person level and group level

Group-level Latent Classes		Person-level Latent Classes			
		Condition 1		Condition 2	
		$P(G=1)$ =0.50	$P(G=2)$ =0.50	$P(G=1)$ =0.30	$P(G=2)$ =0.70
Two Classes	$P(K=1)=0.50$	0.40	0.10	0.20	0.30
	$P(K=2)=0.50$	0.10	0.40	0.10	0.40
Four Classes	$P(K=1)=0.25$	0.05	0.20	0.025	0.225
	$P(K=2)=0.25$	0.10	0.15	0.050	0.200
	$P(K=3)=0.25$	0.15	0.10	0.100	0.150
	$P(K=4)=0.25$	0.20	0.05	0.125	0.125

The estimation of MMIRT models can be time consuming. A typical continuous MMIRT model may take 2 to 3 hours on average to converge, and a simple discrete MMIRT model with only two group-level latent classes requires approximately 30 minutes, on a 3.0 GHz computer with 1GB of RAM. Only 50 replications were conducted to implement the simulation within a manageable time period, while obtaining reasonable stability in the results. In addition, to reduce computing time, the true item difficulty values of all items in one latent class and non-DIF items in another latent class were provided as starting values. Meanwhile, three sets of random starting values were generated for the rest of model parameters. Only the one with the highest log-likelihood value was used for the final stage estimation. The selection of starting values may lead to potential problem of local maximum and reduce the generality of the results somewhat. Furthermore from the pilot study, a continuous model that was unable to converge after 200 iterations could be considered as non-converged. Therefore, the maximum number of iterations for the continuous MMIRT models was limited to 200, which could raise the possibility of non-convergence. Those issues were incorporated in the interpretation and discussion of the simulation findings.

The manipulated factors were fully-crossed, resulting in 48 distinct conditions. Under each condition, true ability levels and class memberships were first sampled from the population distributions specified above. Item-level responses were simulated next with 50 replications. Four estimation models were fitted to the generated dataset, yielding a total of 196×50 distinct model estimations (3200 for the continuous distributions and 6400 for the discrete distributions). Data generation and

model estimation were conducted in *R* 2.14.1 (R Development Core Team, 2011) interfacing with M*plus* 6.12 (Muthén & Muthén, 2010). The item difficulty parameters used for data generation were provided in Table 1 in Appendix A, and the sample M*plus* codes for estimating continuous and discrete MMIRT models were included in Appendix C and Appendix D, respectively.

3.2.3 Evaluation Criteria.

Convergence rate. Given model complexity, non-convergence was expected for some MMIRT models under certain simulated conditions. The convergence-rate within the number of replications was recorded and utilized as an indicator of model performance. The results can be used as empirical guidance for practitioners to properly implement MMIRT methods to data with varying characteristics.

Item parameter recovery. The accuracy of item difficulty recovery was evaluated in terms of average item bias and root mean squared error (RMSE). The item difficulty scale was identified by fixing the distribution of latent ability to a standard normal distribution within each person-level latent class.

Instead of item parameter, the terms of interest was item difficulty difference (Δb) between the two person-level latent classes. Let Δb_{tr} denote the true difficulty difference and $\Delta\hat{b}_r$ the estimated difficulty difference for the *r*th ($r = 1...R$) replication, they are expressed by

$$\Delta b_{tr} = b_{tr,\,focal} - b_{tr,reference}, \tag{3.11}$$

$$\Delta\hat{b}_r = \hat{b}_{r,LC1} - \hat{b}_{r,LC2}. \tag{3.12}$$

Bias and RMSE, therefore are defined with respect to difficulty difference using the following equations, respectively,

$$Bias(\Delta\hat{b}) = \frac{1}{R}\sum_{r=1}^{R}(\left|\Delta\hat{b}_r\right| - \Delta b_{tr}), \tag{3.13}$$

$$RMSE(\Delta\hat{b}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\left|\Delta\hat{b}_r\right| - \Delta b_{tr})^2}. \tag{3.14}$$

Calculating bias and RMSE of item difficulty difference in the above does not require the true and estimated item parameters to be placed on the same scale. These two criteria are averaged across items to provide global recovery of item parameters. To differentiate the influence on different item types, separate analyses were also conducted to DIF items and non-DIF items.

Classification recovery at the person level. Class memberships at the person-level are indicators of group-level latent classes. Whether population distribution at the group-level can be correctly modeled depends on the success in recovery of classification at the person level.

Classification recovery at the person level was evaluated by classification agreement between the true and estimated class memberships. The criterion used is the Cohen's kappa (κ; Cohen, 1960). The Cohen's kappa is a proper measure of agreement between two procedures that measure the same thing. The percentage is computed using a 2×2 class assignment matrix as shown below, in which an individual is assigned to one of the four cells.

		Simulation		
		1	2	Total
Estimation	1	P_{11}	P_{12}	$P_{1.}$
	2	P_{21}	P_{22}	$P_{2.}$
	Total	$P_{.1}$	$P_{.2}$	1

To compute kappa, the observed level of agreement ($P_o = P_{11} + P_{22}$) and the expected level of agreement if the two procedures are totally independent ($P_e = P_{.1}P_{1.} + P_{.2}P_{2.}$) are defined. The corrected kappa is calculated via

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{3.15}$$

Arbitrary guidelines characterize a kappa value below 0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and over 0.80 as almost perfect agreement between the two procedures (Landis & Koch, 1977).

Simulation studies with regard to mixture models frequently encounter the problem of switched class labels. Label switching happens when identified latent classes change meaning during the estimation. Given the fact that class labels are arbitrary, if labels are potentially switched across data sets, aggregating parameter estimates over potential mislabeled classes is undesirable (Tueller, Drotar & Lubke, 2011). Tueller, Drotar and Lubke (2011) further proposed a switched label detection algorithm, where the largest assignment percentage should be achieved on the diagonal of the class assignment matrix. Followed the same logic of the proposed algorithm, an observed level of agreement less than .50 indicates the occurrence of label switching for two person-level mixtures. The computation of P_o and P_e was adjusted according to the value of observed agreement level as shown below,

$$P_o = \begin{cases} P_{11} + P_{22} \\ 1 - (P_{11} + P_{22}) \end{cases} \text{ and } \quad P_e = \begin{cases} P_{.1}P_{1.} + P_{.2}P_{2.} & \text{if } (P_{11} + P_{22}) \geq 0.5 \\ P_{.1}P_{2.} + P_{.2}P_{1.} & \text{if } (P_{11} + P_{22}) < 0.5 \end{cases}. \tag{3.16}$$

Recovery of random effects at the group level. The recovery of random effects at the group level was evaluated by the consistency of person level latent class proportion within group units.

Since random effects at the group level can be specified as continuous random variable or discrete latent classes, it is not feasible to compare directly the true and estimated group level random effects under various model specifications. Due to two person-level mixtures, the proportion of one person-level latent class within groups is sufficient to reflect the change of class probability across groups. The correlation between the true and estimated proportions was utilized to indicate random effect recover at the group level. The group-level latent classes are essentially composed of groups with similar proportions of person-level latent classes, using the correlation also provides a unified criterion for the comparison between the continuous and the discrete approaches. Note that only the absolute values of correlation were kept for analysis to avoid potential label switching problem.

Model selection. The percentage of replications in which the population distribution was correctly identified was used to indicate the power of model fit indices. Both the original indices and the modified versions were compared in terms of how often they selected the true models as the first or the second choice under each simulated condition.

The frequency of correct selection, however, does not provide information on how close the competing models are in terms of fit indices. In the current study, the two models with smallest fit values within each replication were considered as a comparison pair. The two comparison pairs with the highest occurrence probabilities over all simulated conditions were reported. To facilitate the interpretation of difference size, the current study adopted the likelihood ratio approach to compare the

two competing models (Hamaker, et al., 2011). For any information criterion, the value is transformed as

$$IC^{*} = \exp\left(-\frac{1}{2} IC\right), \tag{3.17}$$

and the ratio of the transformed values is computed in the form of

$$\frac{IC^{*}_{1^{st}}}{IC^{*}_{2^{nd}}} = \exp\left(-\frac{1}{2}(IC_{1^{st}} - IC_{2^{nd}})\right), \tag{3.18}$$

where $IC_{1^{st}}$ is fit result for the model with the smallest value and $IC_{2^{nd}}$ is the one with the second smallest value. The interpretation of the ratio is in a similar manner as likelihood ratios (Burnham & Anderson, 2002), such that the first model is said to be "*ratio*" times more likely to be the population model than the second one.

No consensus has been reached on how large the ratio should be so that a model can be considered as the best fitting model with confidence. The current study selected two levels of ratio arbitrarily: 10 and 100, which correspond to an IC difference (ΔIC) of 4.62 and 9.21, respectively, as cutoff values of small and medium ratio size.

The M*plus* output files were read back into *R*, where the evaluation criteria were computed next. Once evaluation criteria were collected, factorial ANOVA was performed to compare model performance on item and classification recovery using PROC GLM (SAS 9.2, SAS Institute, 2009). The four estimation models were dummy coded with the true model specified as the reference model. Only the main effect of estimation models, its two-way and three-way interactions with the four manipulated factors were included in ANOVAs. In addition, eta-squared (η^2) was

employed to present the percentage of variance explained by the main effects and interactions. Only $\eta^2 > 0.05$ was reported, that was an effect explained more than 5% of variance in the outcome variable. Cohen's f^2, defined as $f^2 = \frac{\eta^2}{1-\eta^2}$ was further calculated as a measure of effect size. Cohen suggested that $f^2 \geq 0.15$ is a medium effect size and $f^2 \geq 0.35$ is a large effect size (Cohen, 1988).

Chapter 4: Results

The framework of MMIRT modeling is not new, but the exploration of model function under various conditions still requires extensive research. The current study concentrated on distinguishing between the two MMIRT approaches, continuous and discrete, from a model comparison perspective. The simulation study described in the third chapter intentionally selected four manipulated factors to investigate the performance of six information criteria plus four modified versions in identifying the latent distribution of random effects at the group level. An empirical data analysis was conducted next to determine and illustrate MMIRT model function with regard to model fit criteria.

4.1 Results of Simulation Study

Table 4.1 Variable names of simulation factors in results

Description	Variable
Latent class	
Person-level latent class	PLC
Group-level latent class	GLC
MMIRT models	
Continuous MMIRT model	Cont
Discrete MMIRT model with two group-level latent classes	GLC2
Discrete MMIRT model with three group-level latent classes	GLC3
Discrete MMIRT model with four group-level latent classes	GLC4
Discrete MMIRT model with five group-level latent classes	GLC5
Manipulated factors	
Percentage of DIF items	DIF
Group size	Size
Proportion of person-level latent class	Prop
Ability variance at the group level	Var

For the purpose of clear presentation, model names and manipulated factors were given short abbreviations (listed in Table 4.1) in the following tables and figures.

4.1.1 Non-Convergence Rate.

For every generated dataset, four different MMIRT models, one continuous plus three discrete, were fitted and compared with regard to model parameter recovery, latent class classification recovery as well as model fit statistics. A valid replication should have the four estimation models converged when fitted to the same dataset. Non-convergence occurring with any estimation model would lead to an invalid run. For all simulated conditions, additional iterations were conducted until the number of valid replications reached 50. The detailed non-convergence rates were shown in Appendix B.

Altogether, 18 out of 48 simulated conditions never encountered convergence problems requiring additional iterations, and another 6 conditions were associated with a convergence rate higher than 95%. 11 conditions were found to have a frequency of non-convergence larger than 10, corresponding to a convergence rate less than 80%.

Non-convergence often occurred in discrete MMIRT models, especially when the data-generation model was discrete and more latent classes were extracted at the group level. In contrast to the GLC2 that performed stably throughout all simulated conditions, the GLC4 was often unable to converge even when it was the data-generation model. A close examination revealed that uneven proportion of PLC raised the probability of non-convergence, particularly when coupled with a small percentage of DIF items and small group size.

Similar with the GLC2, the continuous MMIRT model seldom encountered convergence issues with only one exception when the true generation model was GLC2 along with 30% DIF items, group size of 150, uneven PLC proportion and large group-level ability variance. This condition was particularly problematic. Except the true generation model (GLC2), the other three models all had a high non-convergence rate. Consequently, a new set of ability and class parameters were generated for this condition and another 50 fully-converged iterations were conducted.

Table 4.2 Number of free parameters for all fitted models

	Cont	GLC2	GLC3	GLC4	GLC5
Number of Parameters	83	84	86	88	90

4.1.2 Main Effect of Estimation Model.

Model performance was evaluated with respect to two main parts: 1) bias and RMSE for item parameter recovery, and 2) Cohen's corrected kappa and correlation of PLC proportion within groups for classification recovery.

Table 4.3 presented performance of estimation models on the four evaluation criteria. The results showed similar pattern over the three data-generation conditions: larger bias and RMSE between the true and estimated item difficulty difference leads to lower agreement on latent class membership at the person level, which further reduces the recovery of random effects at the group level. It was surprising that among the three data-generation models, only the GLC2 better recovered item parameters and latent class membership. The continuous MMIRT model performed consistently poorly on recovering item parameters and identifying person-level latent

classes, regardless of whether it was used to generate data or not. With increasing number of GLC extracted, however, model performance of discrete MMIRT models became identical with the continuous MMIRT models in terms of item and classification recovery. For instance, the GLC4 model was found to have similar although still slightly better results than the Cont model over the four evaluation criteria. The descriptive statistics of the evaluation criteria for each of the three data-generation models were fully presented in Tables 2a to Table 5c in Appendix A.

Table 4.3 Model performance on evaluation criteria

True Model	Estimation Model							
	Cont		GLC2		GLC3		GLC4	
Cont	M	SD	M	SD	M	SD	M	SD
Bias	0.46	0.53	0.20	0.24	0.28	0.37	0.36	0.45
RMSE	0.50	0.54	0.28	0.30	0.35	0.40	0.42	0.46
Kappa	0.46	0.24	0.52	0.18	0.50	0.20	0.47	0.23
Correlation	0.65	0.31	0.74	0.21	0.70	0.26	0.66	0.30
	Cont		GLC2		GLC3		GLC4	
GLC2	M	SD	M	SD	M	SD	M	SD
Bias	0.31	0.35	0.15	0.18	0.22	0.30	0.24	0.32
RMSE	0.35	0.36	0.21	0.21	0.27	0.31	0.28	0.33
Kappa	0.56	0.20	0.60	0.17	0.58	0.20	0.57	0.20
Correlation	0.77	0.27	0.82	0.23	0.80	0.26	0.79	0.27
	Cont		GLC3		GLC4		GLC5	
GLC4	M	SD	M	SD	M	SD	M	SD
Bias	0.44	0.53	0.32	0.42	0.35	0.47	0.38	0.51
RMSE	0.48	0.53	0.37	0.41	0.40	0.46	0.43	0.50
Kappa	0.49	0.23	0.51	0.21	0.50	0.22	0.49	0.23
Correlation	0.73	0.29	0.77	0.24	0.76	0.26	0.74	0.28

Table 4.4a ANOVA of manipulated factors on the four evaluation criteria (True model: Continuous)

	Item Parameter Bias		Item Parameter RMSE		PLC Classification Recovery		Correlation of PLC Proportion	
Source	F test	η^2	F test	η^2	F test	η^2	F test	η^2
Model	9.97***		8.98***		3.99*		4.11*	
DIF*Model	33.47***	0.26††	48.13***	0.28††	111.83***	0.52††	63.30***	0.36††
Size*Model	5.23**		5.50**		2.34		0.50	
Prop*Model	2.14		1.62		2.76		3.00*	
Var*Model	45.34***	0.36††	67.73***	0.39††	45.09***	0.21†	60.83***	0.35††
DIF*Size*Model	0.86		0.60		2.02		1.38	
DIF*Prop*Model	1.21		1.06		1.00		0.97	
DIF*Var*Model	20.77***	0.16†	30.57***	0.18†	35.46***	0.17†	28.37***	0.16†
Size*Prop*Model	0.72		0.52		0.80		0.83	
Size*Var*Model	3.21*		2.79		3.17*		5.03**	
Prop*Var*Model	1.66		1.35		1.43		2.35	

Note: *, $p < .05$; **, $p < .01$; ***, $p < .001$;
†, medium effect size; ††, large effect size.

Table 4.4b ANOVA of manipulated factors on the four evaluation criteria (True model: GLC2)

	Item Parameter Bias		Item Parameter RMSE		PLC Classification Recovery		Correlation of PLC Proportion	
Source	F test	η^2	F test	η^2	F test	η^2	F test	η^2
Model	4.47*		3.19*		0.58		0.86	
DIF*Model	8.77***	0.11	8.97***	0.11	26.54***	0.27††	12.74***	0.11
Size*Model	1.46		0.67		0.17		3.32*	
Prop*Model	11.37***	0.14†	12.86***	0.15†	38.18***	0.39††	51.23***	0.46††
Var*Model	23.26***	0.28††	24.94***	0.29††	8.27***	0.08	13.92***	0.13
DIF*Size*Model	0.90		0.66		0.08		0.23	
DIF*Prop*Model	3.39*		4.57**		7.26***	0.07	7.78***	0.07
DIF*Var*Model	10.00***	0.12	10.18***	0.12	5.66**		5.13**	
Size*Prop*Model	0.82		0.83		0.48		0.27	
Size*Var*Model	0.41		0.15		0.15		0.68	
Prop*Var*Model	13.17***	0.16†	13.96***	0.16†	5.59**		10.19***	0.09

Note: *, $p < .05$; **, $p < .01$; ***, $p < .001$;
†, medium effect size; ††, large effect size.

Table 4.4c ANOVA of manipulated factors on the four evaluation criteria (True model: GLC4)

	Item Parameter Bias		Item Parameter RMSE		PLC Classification Recovery		Correlation of PLC Proportion	
Source	F test	η^2	F test	η^2	F test	η^2	F test	η^2
Model	0.75		0.78		0.13		0.16	
DIF*Model	5.39**	0.11	6.77**	0.10	10.67***	0.28††	3.66*	0.13
Size*Model	2.64		4.44**	0.06	0.31		0.16	
Prop*Model	4.67**	0.09	5.65**	0.08	5.26**	0.14†	3.90*	0.14†
Var*Model	21.55***	0.42††	34.23***	0.48††	9.73***	0.26††	10.13***	0.35††
DIF*Size*Model	0.08		0.05		0.55		0.55	
DIF*Prop*Model	0.18		0.17		0.25		0.12	
DIF*Var*Model	3.73*	0.07	4.51**	0.06	1.97		1.47	
Size*Prop*Model	0.02		0.02		0.03		0.06	
Size*Var*Model	2.65		4.26*	0.06	0.80		1.08	
Prop*Var*Model	4.64**	0.09	5.12**	0.07	3.23*	0.09	2.60	0.09

Note: *, $p < .05$; **, $p < .01$; ***, $p < .001$;
†, medium effect size; ††, large effect size.

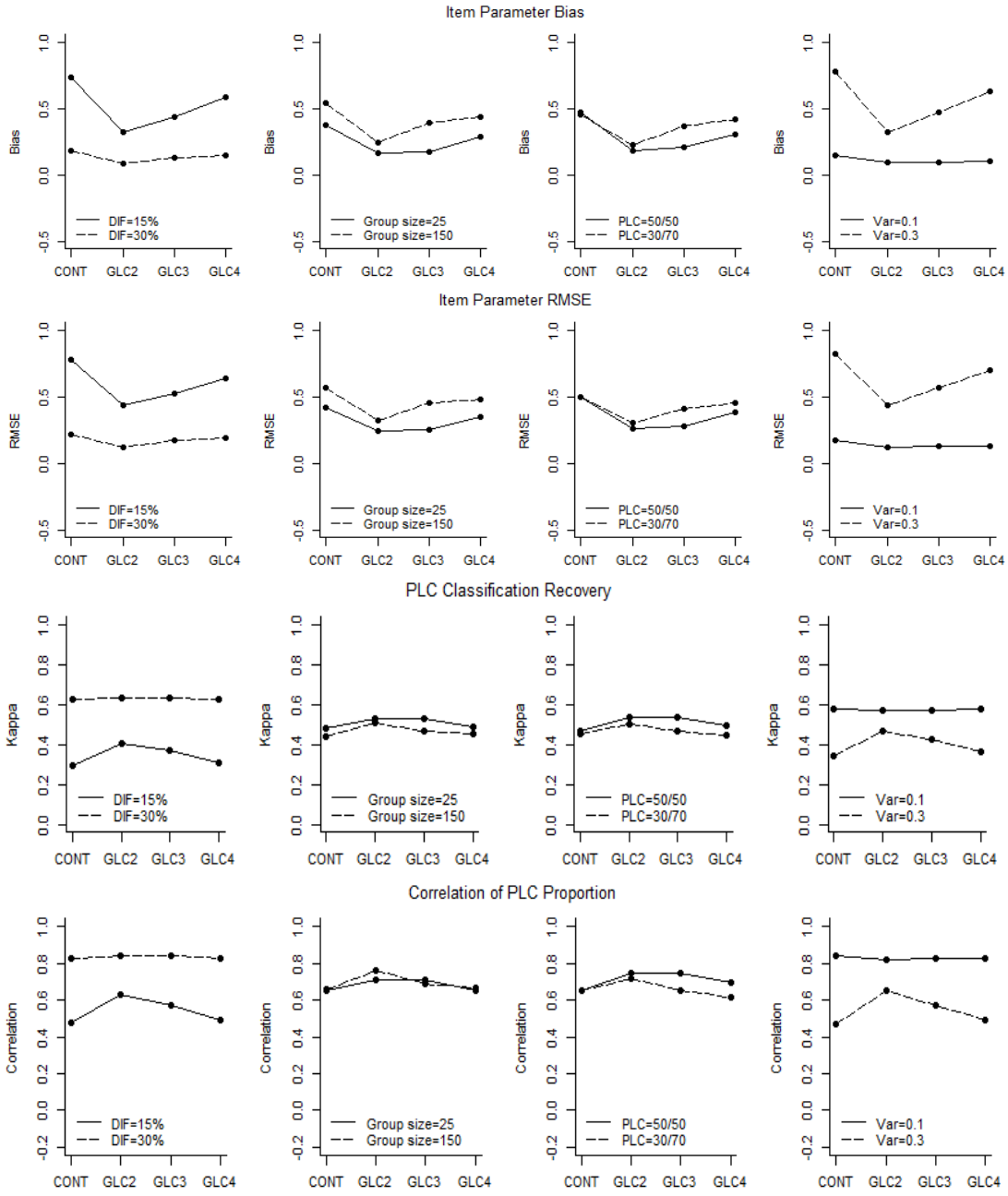

Figure 4.1a. Two-way interactions of manipulated factors and estimation models (True model: Continuous)

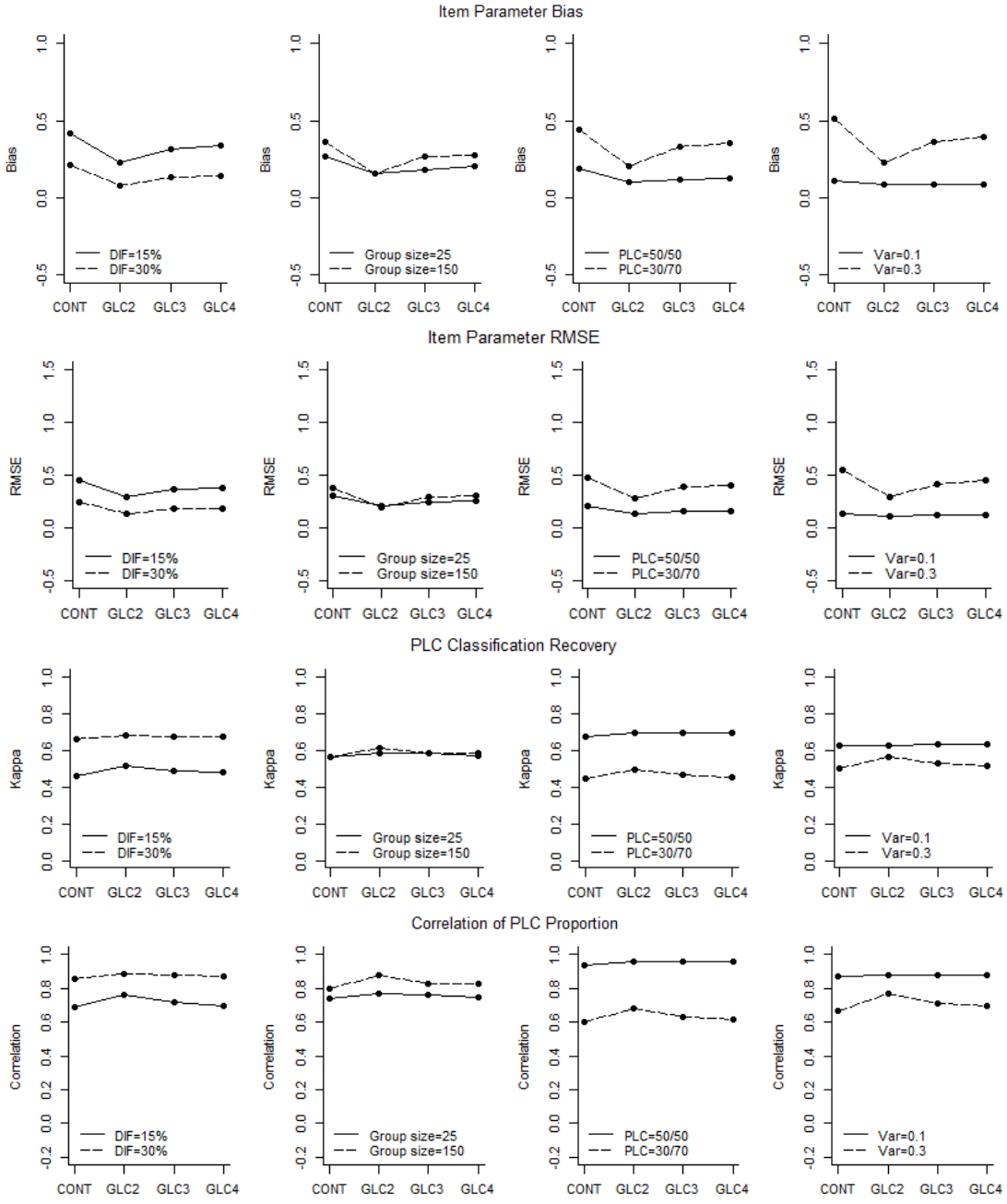

Figure 4.1b. Two-way interactions of manipulated factors and estimation models (True model: GLC2)

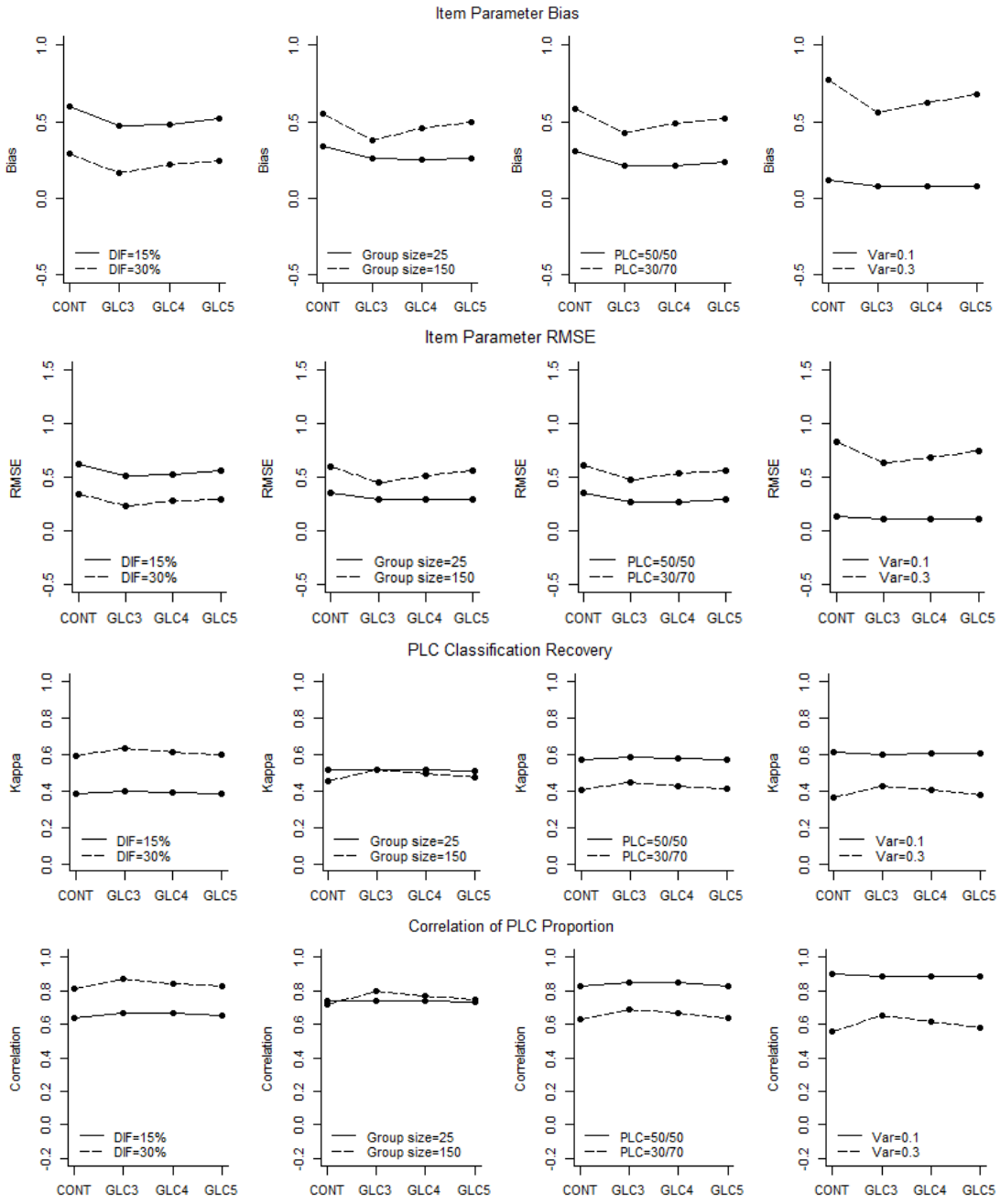

Figure 4.1c. Two-way interactions of manipulated factors and estimation models (True model: GLC4)

The ANOVA results of the evaluation criteria were presented in Tables 4.5a to 4.5c. The main effects of estimation models indicated that significant difference of estimation models only occurred when the data-generation model was the Cont or GLC2. In particular, models differed significantly at $p<.001$ on item parameter recovery for the Cont data. But such effect was less significant under the GLC2. The same pattern was also observed on the classification recovery, where no significant model difference was found under the GLC2. Although F-test was significant, none of the main effect was found to associate with η^2 larger than 5%. For GLC4, estimation models performed identically on both item and classification recovery.

In the following section, the ANOVA results of interactions between the estimation models and manipulated factors on the four evaluation criteria were presented separately. The two-way interactions were depicted by Figures 4.1a to 4.1c to show the trends of performance across the estimation models.

4.1.3 Item Parameter Recovery.

Item parameter recovery is evaluated in terms of item bias and RMSE. Both criteria are defined based on the discrepancy between the estimated and the true item difficulty difference between the two person-level latent classes. According to Equation 3.16, positive bias indicates overestimation of item difficulty difference and negative bias indicates underestimation of difference.

It is noticeable that the pattern of ANOVA results on bias and RMSE are fairly similar over simulation conditions, suggesting a strong relation between these two criteria which is discussed in details in Chapter 5.

True model: Continuous MMIRT model. All two-way interactions except Prop*Model were significant at the $p<.01$ level, but only DIF*Model and Var*Model showed large effect size when interacting with the estimation models. The percentage of DIF items accounted for more than 25% of the variance in bias and RMSE, while the between-group ability variance explained more than 35% of the total variance.

In Figure 4.1a, model performance on item parameter recovery was worse with smaller percentage of DIF items, particularly for the Cont and GLC4 models. Discrete MMIRT models with fewer GLCs (i.e., GLC2 and GLC3) yielded much less bias than the Cont and GLC4 models, when only 15% of items had DIF effect. With increasing number of DIF items introduced in the sample, the four models performed equally well in terms of bias and RMSE. A similar pattern was also observed in the effect of between-group ability variance. Models performed better when between-group ability variance was small (Var=0.1). A large proportion of between-group variance substantially increased estimation bias, and the magnitude of the increment was especially striking in the Cont and GLC4 models. Although group size was found to have some impact on bias and RMSE, with a smaller group size (corresponding to a greater number of groups) leading to a better recovery, the magnitude of difference was not as much as that in DIF item percentage and ability variance. The proportion of PLCs had minimal impact on item recovery.

The only significant three-way interaction, DIF*Var*Model had a medium effect size on item parameter recovery. This interaction accounted for more than 15% of total variance. The trend of this effect on bias was displayed in Figure 4.2. The distinct model difference was only observed under the condition with a small

percentage of DIF items combined with large between-group ability variance. There was minor difference in bias across the four estimation models under the other three conditions. The same pattern was also found on RMSE and is not discussed further.

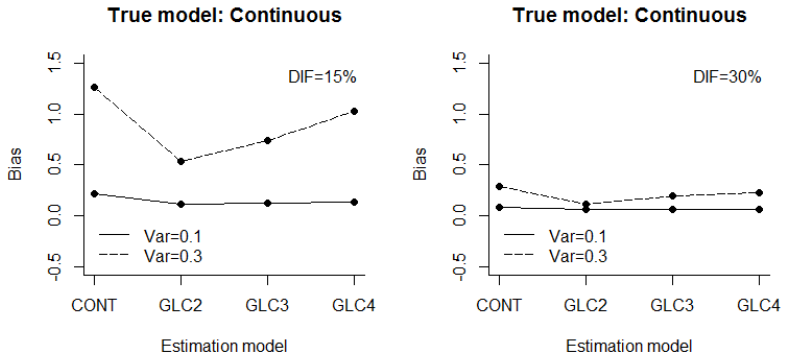

Figure 4.2. Three-way interaction of DIF*Var*Model on item bias

True model: GLC2 MMIRT model. Unlike the Cont model condition, when the GLC2 was the data-generation model, only the interaction between group size and estimation model was found to have no significant effect on item parameter recovery. Among the other three two-way interactions, the Var*Model was still the most important effect, explaining almost 30% of total variance in both bias and RMSE. DIF*Model was still significant at $p<.001$, but its effect size was not as large as in the Cont model. Instead, proportion of PLCs explained around 15% of variance in bias and RMSE, which were medium effect sizes.

Figure 4.1b showed that the effect of between-group variance was similar to what was observed in the Cont condition. A between-group variance of 0.1 yielded much less bias and the difference of bias between the estimation models increased as the variance increased. However, the magnitude of increment in bias was much smaller over the two variance levels than that in the Cont data. For instance, under the condition of Var=0.3 all models had a bias value over 0.5 except the GLC2 when the

true model was the Cont, while no model was associated with a bias value larger than 0.5 in the GLC2 data.

The effect pattern of proportion of PLCs interacting with the estimation models was similar with the ability variance. An even proportion of the two PLCs reduced the estimation bias and RMSE across models. When this proportion became uneven (i.e., 30%:70%), both bias and RMSE increased particularly in the Cont and GLC4 models.

The two proportions of DIF item showed a similar pattern across models where even with 15% of DIF item the bias difference across models was much smaller than that in the Cont data. Furthermore, no group size effect was found on either bias or RMSE.

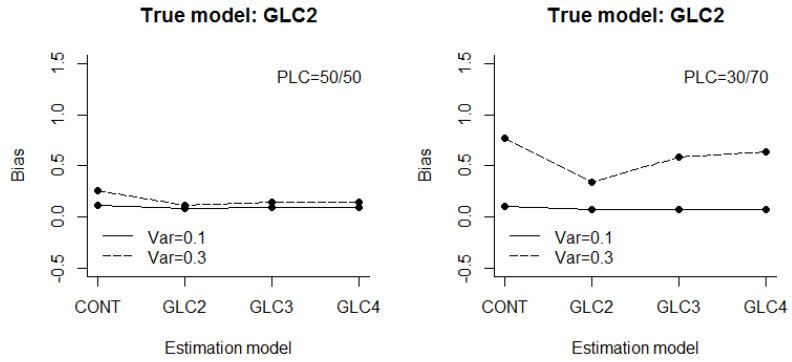

Figure 4.3. Three-way interaction of Prop*Var*Model on item bias

In addition to the significant three-way interaction DIF*Var*Model, Prop*Var*Model was also significant at $p<.001$, and accounted for approximately 16% of the total variance in bias and RMSE, simultaneously. Figure 4.3 showed that when uneven proportion of PLC was coupled with large between-group variance, the bias of GLC2 was much smaller than the other three models. No distinctive difference was observed in bias under the other three simulated conditions. Although

DIF*Prop*Model was significant at $p<.001$, this interaction did not account for more than 5% of variance and was considered to have minor effect on item parameter recovery. The same findings applied to RMSE.

True model: GLC4 MMIRT model. Distinct pattern was found under the GLC4 condition. Although all two-way interactions except Size*Model were significant at $p<.01$ on both bias and RMSE, only Var*Model accounted for larger than 40% of variance.

Again, by looking at Figure 4.1c, the pattern of estimation bias and RMSE was similar to what was found in the Cont data. Consistent with the previous findings, larger between-level variance increased estimation bias and the difference across four estimation models. This time, GLC3 better recovered item parameters. But the average bias of GLC3 under Var=0.3 was over 0.5, larger than that of GLC2 in the Cont data. The performance trend on the other manipulated factors had the same patterns as in the GLC2. One exception was that the group size had significant effect on RMSE, but not on bias. Figure 4.1c showed that with fixed total sample size, more individuals within group (i.e., larger group size and smaller number of groups) increased estimation bias and RMSE.

With regard to three-way interactions, although DIF*Var*Model and Prop*Var*Model were found significant at $p<.05$ on both bias and RMSE, and even Size*Var*Model was significant at $p<.05$ on RMSE, none of them accounted for more than 10% of variance. All three-way interactions had minor impact on item parameter recovery under the GLC4 condition.

Item Type Difference. It is of interest to see how different types of items respond to the manipulated factors. Additional ANOVAs were conducted on both non-DIF and DIF items. Table 4.5 summarized the η^2 and effect size of manipulated factors on item parameter recovery across item types.

Table 4.5 Effect size of manipulated factors on parameter recovery across item types

	Data-generation Models					
	Continuous		GLC2		GLC4	
Source	Non-DIF	DIF	Non-DIF	DIF	Non-DIF	DIF
Bias						
Model		0.13†		0.09		
DIF*Model	0.26††		0.10		0.08	
Size*Model		0.12			0.07	
Prop*Model			0.14†		0.10	
Var*Model	0.36††	0.13	0.27††	0.23†	0.43††	0.16†
DIF*Size*Model				0.10		
DIF*Prop*Model						0.19†
DIF*Var*Model	0.18†		0.12		0.06	
Size*Prop*Model		0.11				
Size*Var*Model					0.07	
Prop*Var*Model			0.16†		0.09	
RMSE						
Model		0.10		0.08		
DIF*Model	0.28††		0.11		0.08	
Size*Model		0.14†			0.07	
Prop*Model			0.16†		0.09	
Var*Model	0.39††	0.19†	0.27††	0.33††	0.46††	0.43††
DIF*Size*Model						
DIF*Prop*Model			0.06			0.12
DIF*Var*Model	0.19†	0.08	0.12			
Size*Prop*Model						
Size*Var*Model					0.06	
Prop*Var*Model			0.17†		0.08	

Note: †, medium effect size; ††, large effect size.

When items were clustered into two groups, most two-way interactions only impacted the non-DIF items significantly. For instance, the effect of Var*Model was

larger on the non-DIF items than on the DIF items across the three distribution conditions. Interestingly, DIF*Model had large effect size only on non-DIF items under the Cont data.

Among all three-way interactions, DIF*Var*Model and Prop*Var*Model accounted for more than 15% of variance in bias and RMSE on non-DIF items under the Cont and GLC2 conditions. For DIF items, the only interaction with medium effect size was DIF*Prop*Model in bias.

4.1.4 Classification recovery.

Classification recovery was evaluated in terms of classification agreement (Cohen's kappa) and correlation of the true and estimated PLC proportions. The ANOVA results of classification recovery were similar to what was observed in item parameter recovery but with several differences.

True model: Continuous MMIRT model. DIF*Model and Var*Model were consistently found to have strong impact on classification recovery. Especially DIF*Model interaction accounted for more than half of variance in kappa, and 36% of variance in correlation. Var*Model had greater influence on correlation with 35% of variance explained, compared to 21% in kappa. The three-way interaction involving DIF and Var was also found significant on both kappa and correlation with a medium effect size. The other two factors, group size and proportion of PLCs were found to have no significant impact on classification recovery though.

The last two rows in Figure 4.1a demonstrated that introducing more DIF items helped increase correct identification of latent class membership. With 30% DIF items, all models were able to get a kappa value over 0.60 and a correlation

larger than 0.80 on average. When the percentage of DIF items reduced to 15% the range of kappa values became 0.20 to 0.40, and the correlation dropped below 0.60. The GLC2 model was the least affected by the change of DIF item proportion, whereas the Cont and GLC4 were the most. The effect of Var*Model shared the same pattern on classification recovery. The estimation models performed equally well with small between-group ability variance. The increase in group-level variance led to dramatic decrease in classification agreement and proportion correlation especially in the Cont and GLC4 models.

Consistent with the results found in item parameter recovery, DIF*Var*Model was the only significant three-way interaction, accounting for more than 15% of total variance in both kappa and correlation. As displayed in Figure 4.4, large model difference was only observed when small percentage of DIF items combined with large between-group variance.

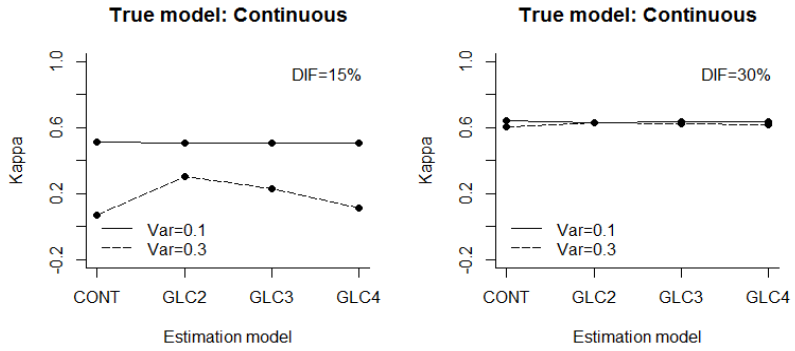

Figure 4.4. Three-way interaction of DIF*Var*Model on kappa

True model: GLC2 MMIRT model. Compared to the large effect on item parameter recovery, Var*Model was found to have small effect size on classification recovery, accounting for 8% and 13% of variance in kappa and correlation, respectively. In contrast, Prop*Model became the most significant effect which

explained around 40% of variance in classification measures. DIF*Model was found to have greater impact on classification agreement (explaining 27% of variance) than on correlation of PLC proportions (explaining only 11% of variance). Two three-way interactions were significant at $p<.001$, including DIF*Prop*Model and Prop*Var*Model. But none of them had medium to large effect size on either kappa or correlation.

In Figure 4.1b, when the data-generation model was the GLC2, an even proportion of PLC in the population tended to greatly improve identification of person-level latent classes, which in turn increased the recovery of random effects at group-level. Such effect was not unique to a particular data-generation model. When this proportion became uneven all the estimation models were affected, although the GLC2 still performed slightly better than the other three models. The trend in DIF*Model and Var*Model was similar with the results found in the Cont data.

True model: GLC4 MMIRT model. When the data-generation model was the GLC4, Var*Model once again was found to be significant at $p<.001$with large effect size on classification recovery. This interaction accounted for 26% and 35% of variance in kappa and correlation, respectively. Prop*Model also had significant effect on classification recovery at $p<.05$, accounting for 14% variance in both criteria. Another important effect, DIF*Model only significantly affected the classification agreement and explained 26% of variance in kappa. But its effect on correlation was much smaller. Even though three out of four two-way interactions were significant, none of the three-way interactions accounted for larger than 10% of the variance in either of the two classification criteria.

Figure 4.1c showed a similar performance pattern as seen in the GLC2 data across the four two-way interactions. One noticeable dissimilarity was on DIF*Model effect, where the GLC2 performed slightly better than the other models even under 30% DIF item condition.

4.1.5 Model Selection.

The frequency of times the data-generation model being selected as the first or the second choice with respect to various information criteria are summarized in Tables 4.7a to 4.7c. The decision to provide frequency instead of percentage was due to the fact that only 50 replications were conducted within each condition. Percentage results can be easily obtained by multiplying the frequency by 2.

The fit indices had no difficulty identifying the correct population model when it was the Cont (as shown in Table 4.6a). Only two simulated conditions were found to cause some problem for the AIC, AICC, nAICC and nssBIC to choose the Cont as the best fitting model. Especially for the condition with 30% DIF item, coupled with the group size of 150, uneven proportion of PLC and small between-group variance, the four indices mentioned above preferred the GLC3 or GLC4 over the Cont, while the other indices still chose the Cont most frequently.

When uneven proportion of PLC was combined with large between-group variance, this condition can reduce the chance of the GLC2 being identified correctly (as shown in Table 4.6b). This effect can be worse when the group size was small, all fit indices pointed to the Cont instead of GLC2 as the best fitting model.

Table 4.6a Frequency of correct model selection (True model: Continuous)

				n = Total Sample Size												*n* = Group Size							
				AIC		AIC3		AICC		CAIC		BIC		ssBIC		nAICC		nCAIC		nBIC		nssBIC	
DIF	Size	Prop	Var	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
15%	25	50/50	0.1	49	1	49	1	49	1	50	0	50	0	50	0	50	0	50	0	50	0	49	1
			0.3	48	2	49	1	48	2	49	1	49	1	49	1	49	1	49	1	49	1	49	1
		30/70	0.1	47	2	47	3	47	2	50	0	50	0	48	2	48	2	49	1	48	2	47	3
			0.3	15	17	23	14	15	18	37	9	37	9	32	12	30	14	35	10	32	12	16	17
	150	50/50	0.1	39	9	44	4	42	6	48	1	48	1	47	2	36	10	45	4	44	5	24	16
			0.3	48	0	48	0	48	0	48	0	48	0	48	0	48	0	48	0	48	0	48	0
		30/70	0.1	48	1	49	0	48	1	49	0	49	0	49	0	39	9	49	0	49	0	24	20
			0.3	48	2	50	0	48	2	50	0	50	0	50	0	47	2	50	0	50	0	44	5
30%	25	50/50	0.1	47	2	49	0	48	1	50	0	50	0	50	0	49	1	50	0	50	0	48	1
			0.3	43	5	48	1	44	4	50	0	50	0	49	1	49	1	50	0	49	1	47	1
		30/70	0.1	49	1	50	0	49	1	50	0	50	0	50	0	50	0	50	0	50	0	50	0
			0.3	47	2	49	0	47	2	50	0	50	0	50	0	50	0	50	0	50	0	47	2
	150	50/50	0.1	45	5	50	0	46	4	50	0	50	0	50	0	41	4	50	0	50	0	15	25
			0.3	48	2	49	1	48	2	50	0	50	0	50	0	43	5	50	0	50	0	30	19
		30/70	0.1	5	12	16	14	5	13	49	1	48	2	39	10	7	7	34	12	21	20	1	5
			0.3	36	11	47	2	36	11	50	0	50	0	50	0	29	17	50	0	50	0	26	18

Table 4.6b Frequency of correct model selection (True model: GLC2)

				n = Total Sample Size												*n* = Group Size							
				AIC		AIC3		AICC		CAIC		BIC		ssBIC		nAICC		nCAIC		nBIC		nssBIC	
DIF	Size	Prop	Var	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
15%	25	50/50	0.1	38	1	39	0	38	1	39	0	39	0	39	0	39	0	39	0	39	0	38	1
			0.3	24	4	27	3	24	4	22	21	24	16	25	10	26	9	25	11	25	10	24	4
		30/70	0.1	10	39	6	43	10	39	0	50	0	50	2	47	2	47	2	48	2	47	10	39
			0.3	2	8	2	8	2	8	0	10	0	10	1	10	1	9	1	9	1	10	2	8
	150	50/50	0.1	47	2	48	1	47	2	50	0	50	0	49	1	46	1	49	1	49	1	42	5
			0.3	17	10	20	11	17	10	21	28	21	28	22	15	16	7	23	11	21	10	13	5
		30/70	0.1	37	7	33	12	37	7	6	41	8	39	22	23	38	4	24	21	31	14	40	2
			0.3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3
30%	25	50/50	0.1	44	5	47	2	44	5	50	0	50	0	49	1	49	1	50	0	49	1	46	3
			0.3	47	3	49	1	47	3	50	0	50	0	50	0	50	0	50	0	50	0	48	2
		30/70	0.1	1	47	0	48	1	47	0	48	0	48	0	48	0	48	0	48	0	48	0	48
			0.3	0	48	0	48	0	48	0	50	0	50	0	49	0	48	0	49	0	49	0	48
	150	50/50	0.1	4	21	9	31	4	22	47	3	45	4	32	16	3	14	23	25	15	29	2	7
			0.3	22	16	31	12	22	16	45	5	45	5	45	3	15	15	43	4	38	7	9	8
		30/70	0.1	26	0	26	0	26	0	21	5	23	3	26	0	25	1	26	0	26	0	23	2
			0.3	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0

Table 4.6c Frequency of correct model selection (True model: GLC4)

				n = Total Sample Size												*n* = Group Size							
				AIC		AIC3		AICC		CAIC		BIC		ssBIC		nAICC		nCAIC		nBIC		nssBIC	
DIF	Size	Prop	Var	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
15%	25	50/50	0.1	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4
			0.3	1	1	1	1	1	1	0	2	0	2	0	2	0	2	0	2	0	2	1	1
		30/70	0.1	1	2	0	1	1	2	0	1	0	1	0	1	0	1	0	1	0	1	1	1
			0.3	2	11	0	4	2	11	0	1	0	1	0	1	0	1	0	1	0	1	0	11
	150	50/50	0.1	1	20	0	17	1	20	0	5	1	4	0	9	6	15	0	14	0	18	13	7
			0.3	0	9	0	8	0	9	0	7	0	10	0	11	1	5	0	10	0	8	2	5
		30/70	0.1	10	30	4	23	10	30	0	4	0	5	0	6	12	32	0	11	0	20	20	27
			0.3	0	9	1	10	0	9	2	31	2	24	2	14	0	10	2	11	1	11	0	10
30%	25	50/50	0.1	2	21	1	4	2	20	0	0	0	0	0	0	0	0	0	0	0	0	2	14
			0.3	1	14	0	3	1	13	0	0	0	0	0	0	0	0	0	0	0	0	1	9
		30/70	0.1	16	25	7	26	15	25	0	20	0	20	0	20	0	21	0	20	0	20	13	23
			0.3	12	21	4	17	11	22	0	12	0	12	0	14	0	15	0	14	0	14	9	19
	150	50/50	0.1	34	11	17	32	33	13	0	30	0	36	2	45	38	6	2	47	6	43	39	10
			0.3	29	11	19	23	29	11	1	21	1	23	2	35	31	10	2	37	5	36	32	11
		30/70	0.1	14	29	8	20	13	30	0	4	1	3	1	8	17	25	1	14	5	14	20	24
			0.3	1	15	1	16	1	15	1	20	2	19	2	17	3	12	2	15	1	16	3	11

The fairly low selection rate in the GLC4 supported the earlier argument that when more GLCs were introduced in the discrete model, to discriminate it from a continuous model became increasingly difficult. The selection pattern suggested that with 30% of DIF items, the AIC, AIC3, AICC, nAICC and nssBIC still had a better chance to correctly identify the population model. However, when percentage of DIF items interacted with the other mixture features, it became harder to identify the GLC4 even for those five indices.

Figure 4.5 summarized the overall percentage of model selection for the four estimation models across a total of 16×50=800 replications under each of the three data-generation models. The four adjacent histogram bars present the results for the four estimation models. In line with the order used previously, when the population model was the Cont or GLC2, the four models from left to right were Cont, GLC2, GLC3 and GLC4; when it was GLC4, the four models were Cont, GLC3, GLC4 and GLC5. The bars with darker color represent the model was chosen as the best fitting model; lighter color bars represent that it was the second best fitting model.

The selection pattern in Figure 4.5 indicated that the Cont model always had the highest chance to be selected as the best fitting model when the decision was made based on the CAIC, BIC and ssBIC, nCAIC as well as nBIC. Comparing to the fairly low selection rate of discrete MMIRT models under the Cont data, this rate increased based on the AIC, AIC3, AICC, nAICC and nssBIC when the data-generation model was GLC4. For severely discrete distribution, such as GLC2, the AIC, AIC3, AICC, and nssBIC favored the GLC2 over the Cont model, although the selection rate of the Cont was only slightly lower.

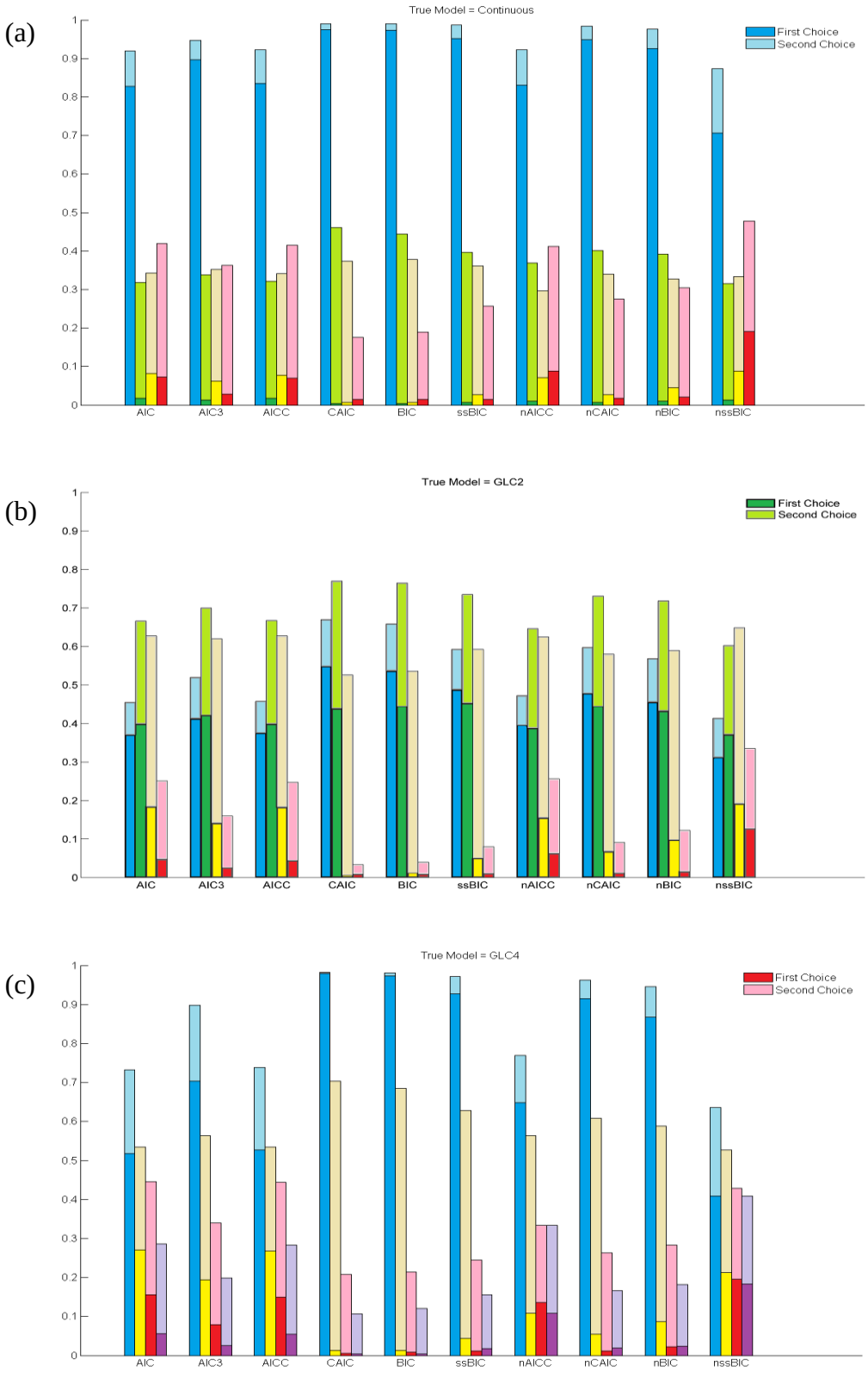

Figure 4.5. Overall percentage of model selection across simulated conditions

Model Comparison. To examine differences of competing models, Tables 4.8a to 4.8c summarized the two comparison pairs with the highest occurrence probabilities over all simulated conditions. The IC differences between the first and second selections were provided as a measure of ratio size.

For the Cont data, all fit indices except AIC most often selected the Cont as the best fitting model and the GLC2 as the second best model (as shown in Table 4.7a). More than 80% of comparisons between the two model were associated with a ΔIC larger than 9, indicating that the support in the data was about 100 times larger for the Cont than for the GLC2. In contrast, the AIC made decision between the Cont and GLC4 more often, but only 38% of the comparisons had a ΔIC larger than 9.

The comparison pair with second highest occurrence probability differed across criteria. More precisely, the AIC3, AICC, nAICC and nssBIC chose the Cont and GLC4 as the second most frequent pair, while the CAIC, BIC, ssBIC, nCAIC and nBIC chose the Cont and GLC3. It can be seen from Table 4.7a that when the decision was made between Cont and GLC2 or GLC3 the magnitude of ΔIC was on average larger than that between Cont and GLC4. Alternatively stated, it is easier to discriminate between Cont and the discrete MMIRT models with fewer GLCs.

The similar pattern can also be found under the GLC4 condition (as shown in Table 4.7c). The comparison pair with highest percentage was between Cont and GLC3. The AIC, AICC and nAICC reached the same selection on the second comparison pair, Cont and GLC5; whereas the rest indices except nssBIC chose between Cont and GLC4. The selection of nssBIC was different from the others. For the nssBIC, GLC4 and GLC5 were compared in 15% of the total iterations with the

ΔIC smaller than 4.21, 94% of times. The fact that all information criteria favored the Cont over GLC4 when the latter was the data-generation model highlights the difficulty in discriminating between the continuous MMIRT model and the discrete models with large number of GLCs.

When the discrete distribution was the GLC2, most fit indices can differentiate between the Cont and the discrete models. It can be seen in Table 4.7b that the decision was made most often between the GLC2 and GLC3 by all fit indices. The size of ΔIC, however differed across indices. The CAIC, BIC, ssBIC, nCAIC and nBIC frequently showed a medium size of ΔIC, compared to the small ΔIC in the AIC, AIC3, AICC and nssBIC. Moreover, the comparison pair, the Cont and GLC2, had the second highest percentage across fit indices.

To better describe the index function, the ten information criteria were categorized into three groups characterized by consistent selection patterns over the three data-generation models. The AIC, AICC and AIC3 often reached converging results with regard to competing models and ΔIC size. Those three indices are called Type-A indices. The CAIC, BIC, their modified versions and ssBIC performed similarly, and are grouped into Type-B indices. The function of the nAICC and nssBIC was unlike the previous two types, and they are called Type-C indices.

Table 4.7a Model comparison between the first and second choice (True model: Continuous)

	First Pair					Second Pair				
	Selection		ΔIC			Selection		ΔIC		
Index	Pair	%	Range	>4.61 %	>9.21 %	Pair	%	Range	>4.61 %	>9.21 %
AIC	CONT-GLC4	30	(0.22, 258.16)	72	38	CONT-GLC2	30	(0.02,1044.72)	87	77
AIC3	CONT-GLC2	32	(0.08,1045.72)	88	75	CONT-GLC4	30	(0.26, 263.16)	89	67
AICC	CONT-GLC2	30	(0.08,1044.78)	87	76	CONT-GLC4	30	(0.05, 258.45)	76	40
CAIC	CONT-GLC2	46	(2.68,1052.42)	99	94	CONT-GLC3	37	(2.50, 689.74)	99	99
BIC	CONT-GLC2	44	(1.68,1051.42)	97	91	CONT-GLC3	37	(1.28, 686.74)	99	96
ssBIC	CONT-GLC2	38	(0.52,1048.24)	95	82	CONT-GLC3	33	(0.15, 677.21)	93	84
nAICC	CONT-GLC2	35	(0.06,1044.30)	93	80	CONT-GLC4	28	(0.08, 255.40)	70	52
nCAIC	CONT-GLC2	39	(0.17,1047.41)	96	84	CONT-GLC3	31	(0.15, 680.08)	91	79
nBIC	CONT-GLC2	37	(0.80,1046.41)	95	82	CONT-GLC3	28	(0.19, 677.08)	90	77
nssBIC	CONT-GLC2	29	(0.13,1043.28)	88	78	CONT-GLC4	21	(0.14, 250.96)	65	43

Table 4.7b Model comparison between the first and second choice (True model: GLC2)

	First Pair					Second Pair				
	Selection		ΔIC			Selection		ΔIC		
Index	Pair	%	Range	>4.61 %	>9.21 %	Pair	%	Range	>4.61 %	>9.21 %
AIC	GLC2-GLC3	34	(0.04, 4.64)	0	0	CONT-GLC2	19	(0.04,596.48)	71	66
AIC3	GLC2-GLC3	36	(0.40, 6.64)	77	0	CONT-GLC2	21	(0.12,597.48)	68	63
AICC	GLC2-GLC3	34	(0.12, 4.76)	1	0	CONT-GLC2	19	(0.10,596.54)	71	66
CAIC	CONT-GLC2	33	(0.22,604.18)	87	66	GLC2-GLC3	33	(0.18, 19.88)	98	91
BIC	GLC2-GLC3	34	(0.02, 17.88)	97	88	CONT-GLC2	32	(0.02,603.18)	86	61
ssBIC	GLC2-GLC3	38	(0.20, 11.60)	87	67	CONT-GLC2	26	(0.06,600.00)	75	57
nAICC	GLC2-GLC3	35	(0.71, 10.31)	53	35	CONT-GLC2	21	(0.06,596.06)	72	63
nCAIC	GLC2-GLC3	36	(0.20, 13.50)	87	68	CONT-GLC2	25	(0.07,599.17)	83	60
nBIC	GLC2-GLC3	36	(0.02, 11.52)	86	41	CONT-GLC2	24	(0.01,598.17)	78	60
nssBIC	GLC2-GLC3	34	(0.04, 5.26)	27	0	CONT-GLC2	19	(0.10,595.04)	72	68

Table 4.7c Model comparison between the first and second choice (True model: GLC4)

	First Pair					Second Pair				
	Selection		ΔIC			Selection		ΔIC		
Index	Pair	%	Range	>4.61 %	>9.21 %	Pair	%	Range	>4.61 %	>9.21 %
AIC	CONT-GLC3	24	(0.08,716.78)	58	38	CONT-GLC5	15	(0.52,663.98)	70	42
AIC3	CONT-GLC3	36	(0.06,719.78)	67	33	CONT-GLC4	18	(0.01,670.34)	66	35
AICC	CONT-GLC3	25	(0.01,716.95)	57	37	CONT-GLC5	15	(0.26,664.40)	72	44
CAIC	CONT-GLC3	69	(2.18,739.88)	100	96	CONT-GLC4	19	(2.10,703.84)	99	98
BIC	CONT-GLC3	67	(1.66,736.88)	98	95	CONT-GLC4	19	(1.36,698.84)	99	98
ssBIC	CONT-GLC3	58	(0.25,727.35)	88	66	CONT-GLC4	22	(4.69,682.95)	100	92
nAICC	CONT-GLC3	43	(0.08,715.31)	85	57	CONT-GLC5	12	(0.21,659.71)	59	38
nCAIC	CONT-GLC3	55	(0.27,724.85)	90	76	CONT-GLC4	23	(0.52,678.78)	94	74
nBIC	CONT-GLC3	50	(0.01,721.85)	87	65	CONT-GLC4	22	(0.46,673.78)	77	57
nssBIC	CONT-GLC3	27	(0.01,712.46)	58	36	GLC4-GLC5	15	(0.08, 6.46)	6	0

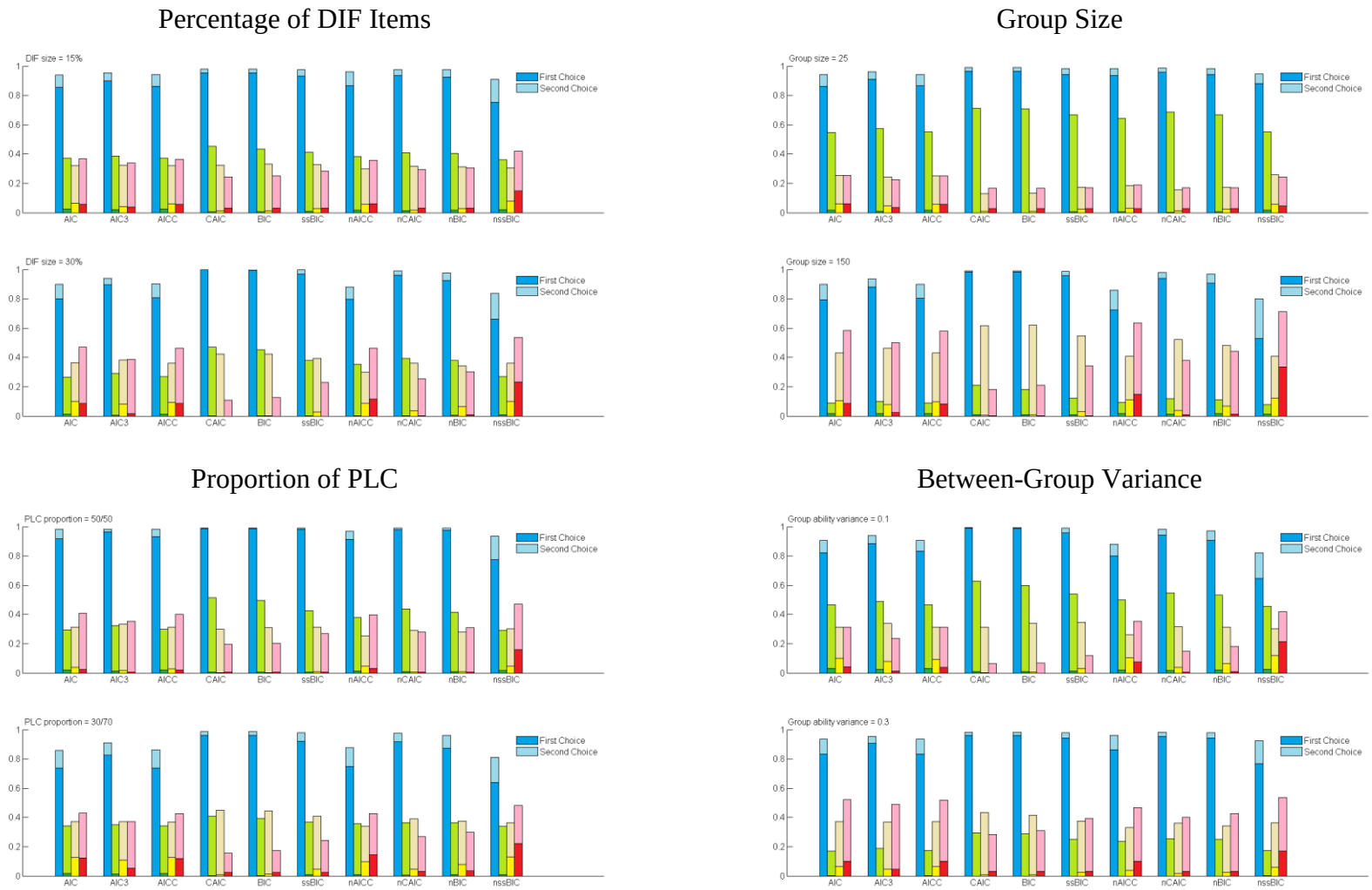

Figure 4.6a. Main effects of manipulated factors on model selection (True model: Continuous)

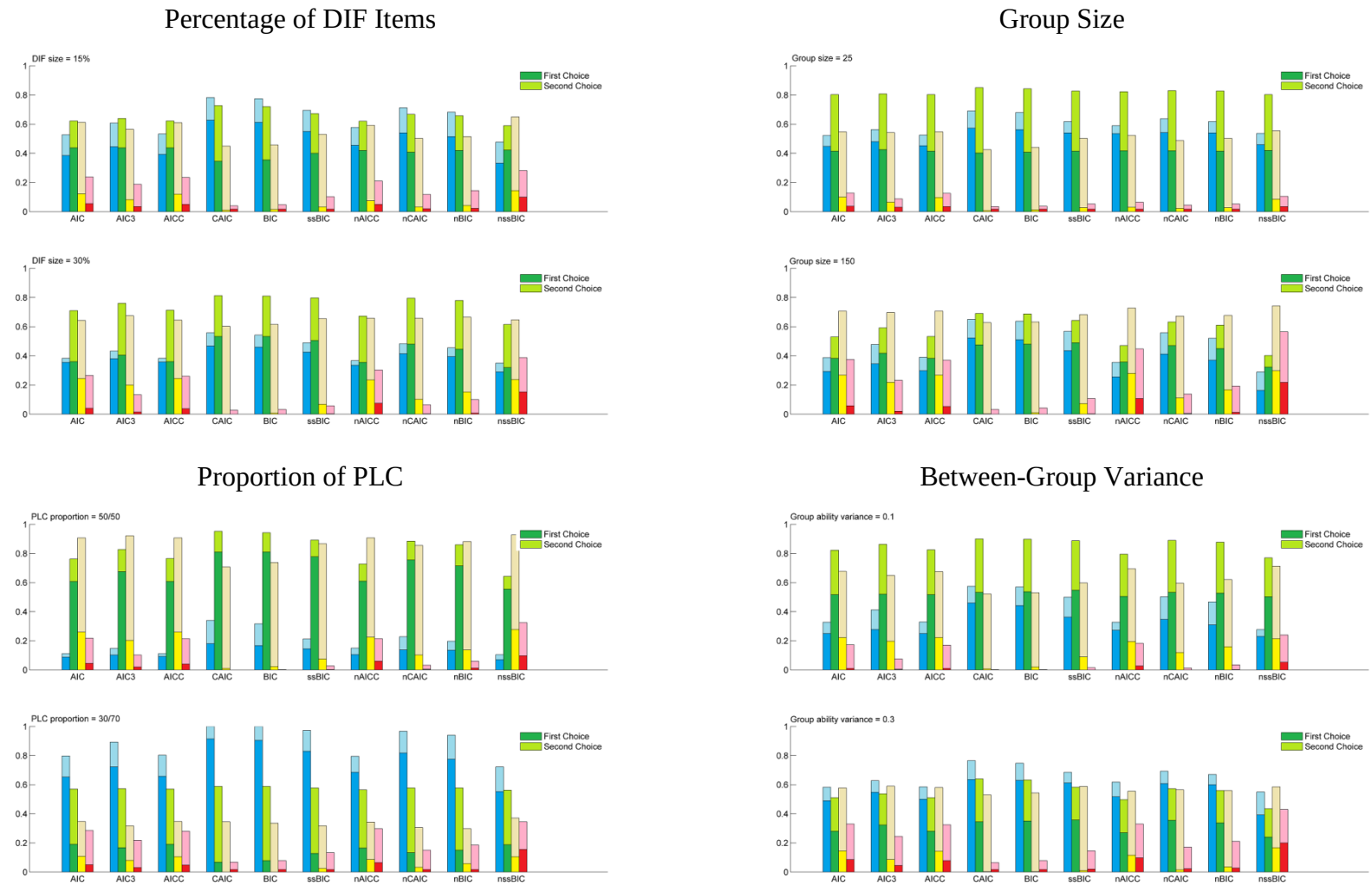

Figure 4.6b. Main effects of manipulated factors on model selection (True model: GLC2)

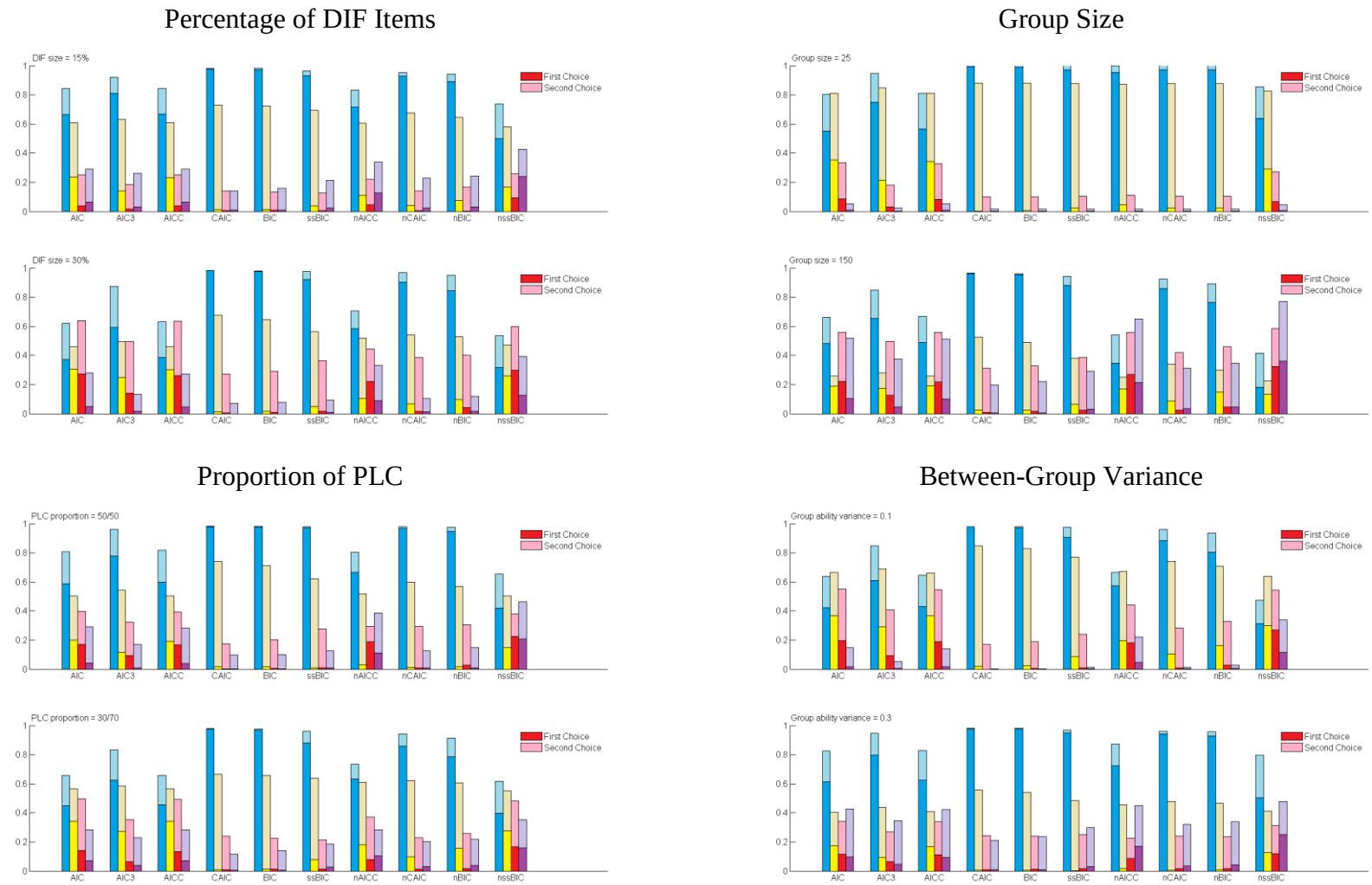

Figure 4.6c. Main effects of manipulated factors on model selection (True model: GLC4)

Effect of Manipulated Factors. Model identification was not only affected by the population distribution, but also by the other properties of the datasets. The main effects of the four manipulated factors on model selection were presented in Figures 4.6a to 4.6c.

As indicated in the previous section, percentage of DIF items and between-group variance were the two most important factors that impacted model performance. Pertaining to model select, larger number of DIF items seemed to increase the probability of the true model being selected as either the first or the second choice. This trend was also observed in the Type-B indices under the GLC2 data. Increasing between-group variance consistently reduced the chance of selecting the GLC2 or GLC3 models as the best or second best model, regardless of the population model. In contrast, the GLC4 or GLC5 models were more likely to be selected with large between-group variance. Such a tendency was not unique to any type of indices, and can be a problem especially when the data-generation model was the GLC2.

With a small group size of 25 in the sample, corresponding to a number of 240 groups, the ten indices showed identical selection patterns. Once the number of groups dropped to 40, the three groups of indices responded dissimilarly to this change. Discrete MMIRT models with more than three GLCs were generally favored over the GLC2. In particular, the Type-A and Type-C indices chose GLC4 more often when the data-generation model was the Cont or GLC4, while the Type-B indices chose the GLC3. The change of group size was expected to substantially affect indices with sample size in their equations. The results partially supported this

statement. For instance, although the nCAIC and nBIC usually showed identical pattern with the CAIC and BIC, the decrease of number of groups caused these two indices to favor the GLC3 and GLC4, rather than the GLC3 as with the application of the CAIC and BIC.

The proportion of PLC was found to have no substantial effect on model selection under the Cont or GLC4 conditions. But it can significantly impact performance of fit indices under the GLC2 condition. As shown in Figure 4.6b, when the proportion of two PLCs changed from even to uneven, the best fitting model switched from the GLC2, the data-generation model, to the Cont. This trend was shared by the ten indices.

Taken together, the above results suggest that a discrete model with fewer than two or three GLCs can recover item parameter and classification fairly well even when it was not the population model. The Cont performed poorly on parameter recovery however, it has been selected more frequently as the best fitting model by the ten fit indices. These findings raise the concern of the application of MMIRT in a real dataset where the population parameters are unknown and fit indices are heavily relied on to make model selection decision.

4.2 Empirical Illustration: MSA Math

The performance of the proposed MMIRT models was further illustrated using an empirical data set. The sample was selected from the 2009-2010 academic year Maryland School Assessment (MSA) math achievement test for the 6^{th} grade. This test included 62 operational items, of which 7 were polytomously scored. For the

Rasch-based MMIRT models proposed in the current study, only the 55 dichotomously scored items were kept for further analysis.

The item-level responses of students from a large suburban county in Maryland were included along with the unique identification numbers of students, and their schools and teachers. To ensure sufficient variation within schools and teachers, teachers with fewer than 15 students were deleted from the sample. The final sample included 3,197 students, 93 teachers and 28 schools. The average classroom size was 34.38 ($SD = 12.89$, range = (15, 66)), and the average school size was 114.18 ($SD = 48.93$, range = (18, 241)).

The empirical data analyses were composed of three parts: 1) a set of conventional mixture Rasch models were fitted to determine the number of latent classes at the student level (SLC); 2) once the number of SLC was determined, a continuous MMIRT model and a set of discrete MMIRT models with various number of GLCs were further fitted with the teachers as grouping variable; 3) the same MMIRT models were also fitted but with the schools as grouping variable.

4.2.1 Mixture Rasch Models.

The first step ignored the clustering of students in classrooms or schools. Table 4.8 presented the model fit indices for mixture Rasch models with two (SLC2) to seven (SLC7) latent classes. The AIC, AIC3, AICC all favored the most complex model (SLC7). The CAIC and BIC both agreed that SLC4 was the optimal solution, while the ssBIC had the smallest fit value on SLC5.

Since there was no conclusive evidence to support the use of any of the six fit indices for traditional mixture models, other information such as entropy and factor

means within latent classes were also incorporated to make final decision. It was seen that entropy of the SLC4 was higher than the other models except the SLC2. A closer look at the classification results showed that the students were classified with regard to their latent ability level. For instance, a two-class solution divided the students into a high-ability and a low-ability group. All item difficulty parameters were significantly lower in the high-ability group than in the low-ability group. Adding more classes further separated the large ability classes into smaller ones. The decrease of entropy value indicated the difficulty to distinguish between those resulting classes. Also considering substantive interpretation of classes, as well as further characterizing group units, too many lower-level latent classes can raise potential problem. Therefore, the SLC4 with relatively large entropy was considered as the best fitting model. Since the SLC5 also showed fairly close results on the BIC and CAIC, this solution was included to compare with the SLC4.

Table 4.8 Fit indices for mixture Rasch Models

	SLC2	SLC3	SLC4	SLC5	SLC6	SLC7
LL	-90792	-90316	-89860	-89633	-89496	-89383
p	113	170	227	284	341	398
Entropy	0.85	0.76	0.79	0.76	0.77	0.75
AIC	181810.89	180972.34	180173.32	179833.46	179673.85	**179561.21**
AIC3	181923.89	181142.34	180400.32	180117.46	180014.85	**179959.21**
AICC	181819.25	180991.55	180208.18	179889.05	179755.55	**179674.72**
CAIC	182609.80	182174.23	**181778.20**	181841.33	182084.71	182375.05
BIC	182496.80	182004.23	**181551.20**	181557.33	181743.71	181977.05
ssBIC	182137.75	181464.07	180829.92	**180654.94**	180660.21	180712.44

Note: *LL* is the log-likelihood value, *p* is the number of model parameters. Results with the smallest value for a particular fit index are in bold.

Table 4.9a Fit indices for teacher-level MMIRT models

	Student-level Latent Classes									
	SLC4					SLC5				
Teacher-Level LC	Cont	GLC2	GLC3	GLC4	GLC5	Cont	GLC2	GLC3	GLC4	GLC5
p	230	231	235	239	243	288	289	294	299	304
Entropy	0.82	0.86	0.89	0.89	0.90	0.79	0.84	0.90	0.90	0.91
Model Selection										
AIC	178732.95	179249.91	178880.48	178769.60	**178688.77**	**178385.75**	178890.90	178904.11	178859.52	178654.32
AIC3	178962.95	179480.91	179115.48	179008.60	**178931.77**	**178673.75**	179179.90	179198.11	179158.52	178958.32
AICC	178768.78	179286.06	178917.94	178808.39	**178728.93**	**178442.99**	178948.56	178963.88	178921.45	178718.44
CAIC	**180359.05**	180883.07	180541.92	180459.32	180406.77	**180421.90**	180934.12	180982.68	180973.44	180803.59
BIC	**180129.05**	180652.07	180306.92	180220.32	180163.77	**180133.90**	180645.12	180688.68	180674.44	180499.59
ssBIC	**179398.24**	179918.08	179560.23	179460.91	179391.66	**179218.80**	179726.85	179754.51	179724.39	179533.65
nAICC	177962.95	178478.80	178104.82	177989.19	**177903.44**	**177536.44**	178040.04	178045.39	177992.86	177779.60
nCAIC	**179545.45**	180065.94	179710.64	179613.89	179547.19	**179403.14**	179911.82	179942.69	179915.77	179728.23
nBIC	179315.45	179834.94	179475.64	179374.89	**179304.19**	**179115.14**	179622.82	179648.69	179616.77	179424.23
nssBIC	178589.39	179105.72	178733.80	178620.42	**178537.09**	**178205.99**	178710.51	178720.60	178672.89	178464.57

Note: *p* is the number of model parameters. Results with the smallest value for a particular fit index are in bold.

Table 4.9b Fit indices for school-level MMIRT models

	Student-level Latent Classes									
	SLC = 4					SLC = 5				
School-Level LC	Cont	GLC2	GLC3	GLC4	GLC5	Cont	GLC2	GLC3	GLC4	GLC5
p	230	231	235	239	243	288	289	294	299	304
Entropy	0.80	0.87	0.89	0.90	0.91	0.80	0.84	0.90	0.91	0.88
Model Selection										
AIC	179547.43	179627.26	179567.15	179559.81	**179457.56**	179352.27	179289.95	179652.33	179615.58	**179225.02**
AIC3	179777.43	179858.26	179802.15	179798.81	**179700.56**	179640.27	179578.95	179946.33	179914.58	**179529.02**
AICC	179583.25	179663.41	179604.61	179598.60	**179497.71**	179409.51	179347.61	179712.10	179677.50	**179289.14**
CAIC	**181173.52**	181260.42	181228.59	181249.53	181175.56	181388.42	**181333.17**	181730.90	181729.50	181374.29
BIC	180943.52	181029.42	180993.59	181010.53	**180932.56**	181100.42	**181044.17**	181436.90	181430.50	181070.29
ssBIC	180212.71	180295.43	180246.89	180251.12	**180160.44**	180185.32	180125.89	180502.73	180480.45	**180104.35**
nAICC	178994.09	179072.19	179005.12	178990.76	**178881.44**	178693.98	178629.80	178982.82	178936.69	**178536.71**
nCAIC	180083.83	180166.00	180115.21	180117.20	**180024.28**	180023.94	179963.96	180338.00	180312.91	**179934.01**
nBIC	179853.83	179935.00	179880.21	179878.20	**179781.28**	179735.94	179674.96	180044.00	180013.91	**179630.01**
nssBIC	179138.75	179216.80	179149.58	179135.14	**179025.78**	178840.53	178776.44	179129.93	179084.30	**178684.85**

Note: *p* is the number of model parameters. Results with the smallest value for a particular fit index are in bold.

4.2.2 Teacher-level MMIRT Models.

Five MMIRT models were fitted to the sample data with teachers as clustering units. Followed the abbreviations used previously, the five models included the Cont model and four discrete models with two to five GLCs.

The results of fit indices for teacher-level models were summarized in Table 4.9a. The large decline in the fit indices with the addition of group-level random effects provided substantial evidence to support the use of MMIRT models to account for the nested structure of the data. When choosing the SLC4 as the base-model at lower level, the GLC5 was chosen as the best fitting model by the Type-A and Type-C indices. Although Type-B indices favored the Cont, the increase in fit values was small in the GLC5. Recalling the findings in the simulation study, such a response pattern among the fit indices was more likely to suggest a discrete distribution in the population. Therefore, the final decision was to select the GLC5 instead of Cont as the optimal solution for the teacher-level model.

In contrast, if the SLC5 was chosen for the student level, the ten fit indices all pointed to the Cont as the best fitting model. Note that the fit value of most indices was smaller for the Cont combined with SLC5 than that for the GLC5 with SLC4. This result raised the question regarding comprehensive consideration of model selection at both lower and higher level in MMIRT models. For the final model, the GLC5 combined with SLC4, the classification results were shown in Table 4.10. Figure 4.6(a) presented the proportions of SLCs within the identified five teacher-level latent classes (Tch_LC). SLCs were named based on their mean latent ability levels. The Tch_LCs were put in order to reflect the proportion of low-ability students

from high to low. The number in parenthesis indicated how many teachers had been classified into a particular Tch_LC. The Tch_LCs were characterized with distinctive distribution patterns of SLCs. For instance, Tch_LC1 was comprised of teachers with a large number of low-ability students (≥70%) and quite a few high-ability students (<5%), whereas Tch_LC5 was featured with the domination of high-ability students and the absence of low-ability students.

Table 4.10 Classification results of empirical sample data

		SLC			
	n	Low	Moderate-low	Moderate-high	High
Tch_LC1	27	367	14	107	12
Tch_LC2	19	64	95	110	6
Tch_LC3	8	129	33	545	167
Tch_LC4	25	32	343	233	418
Tch_LC5	14	1	46	45	430
Sch_LC1	13	675	192	340	59
Sch_LC2	2	61	83	22	0
Sch_LC3	3	70	9	196	90
Sch_LC4	7	182	181	407	348
Sch_LC5	3	0	52	95	135

4.2.3 School-level MMIRT Models.

Although the number of schools was much smaller than the number of teachers in the sample data, the model selection showed similar results based on the ten fit indices (as shown in Table 4.9b). But this time, when combined with the SLC4, the ten fit indices consistently favored the GLC5 over Cont. The strong discreteness at higher level also was reflected in the selection with the SLC5 as the base model. None of the fit indices chose Cont as the best fitting model. Either the GLC2 or

GLC5 were of consideration. However, a careful check revealed that the GLC3, GLC4 and GLC5 all contained one GLC with zero cases. Therefore, if SLC5 was chosen as a lower level solution, the GLC2 model should be considered at the school level.

The classification results of the GLC5 solution for school-level units were provided in Table 4.10, and the proportions of SLC within the identified five school-level latent classes (Sch_LC) were depicted in Figure 4.7(b). Note that 13 out of 28 schools were classified into one latent class with relatively large proportion of moderate to low ability students (nearly 70%), while this portion was less than20% for the three schools within the last Sch_LC.

For comparison purposes, the solution of the GLC2 combined with SLC5 was also presented in Figure 4.7(c). A medium-ability SLC emerged from the five-SLC solution. The first Sch_LC was comprised of more moderate to low ability students, compared to the second GLC with relatively larger proportion of medium to high ability students. However, none of the five SLCs dominated either one of the GLCs. Further examination indicated that the combination of the first two GLCs under the GLC5 solution formed the first GLC under the GLC2 solution, while the last three comprised the second one. Only one school switched its class membership across the two solutions.

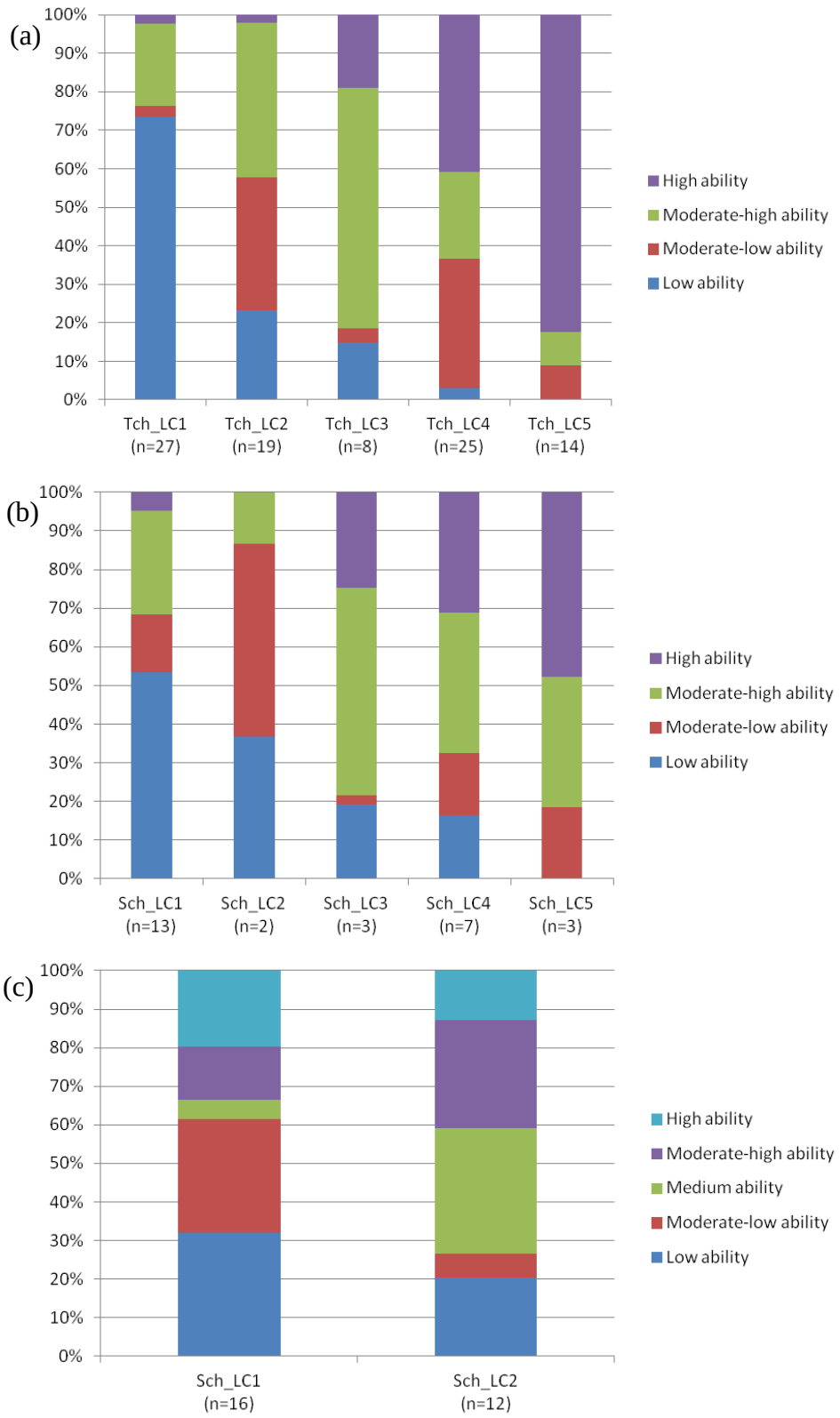

Figure 4.7. Discrete MMIRT solutions at the teacher and school level

Chapter 5: Discussion

Given the emphasis in education research to understand and qualify the effect of teachers and schools on student learning, the current study introduced the MMIRT framework. This framework is capable of capturing population heterogeneity at a contextual-level. As a new approach, the gap between the theoretical discussion of model properties and model performance in empirical analysis remains sizable. In addition to introducing the two MMIRT approaches, the main focus of the current study was to distinguish continuous and discrete MMIRT models under a variety of conditions using a model comparison perspective. A simulation study and an empirical analysis were conducted to evaluate model performance of a set of MMIRT models. The major results are summarized and discussed in this chapter.

5.1 Discussion of Simulation Findings

Model performance of MMIRT models was evaluated in terms of item parameter and classification recovery. The four evaluation criteria, bias and RMSE for item parameter recovery and Cohen's kappa and correlation of PLC proportion within groups for classification recovery, yielded coherent conclusions with respect to the model performance. A brief summary of findings is listed below:

1) Discrete MMIRT models with smaller numbers of GLCs extracted performed consistently better on parameter and classification recovery than the Cont and discrete models with more than four GLCs, regardless of the data-generation models.

2) Four fitted models differed significantly over simulated conditions only when the data-generation model was the Cont. Marginal or insignificant model differences were observed for either parameter and classification recovery under the GLC2 or GLC4 conditions.
3) Throughout the simulated conditions, the Cont model can be correctly identified by most fit indices. Most often, the Cont also was chosen as the best fitting model when the data-generation model was in fact the GLC4. The GLC2 was favored over the Cont by the AIC, AIC3, AICC, nAIC and nssBIC when the former one was used to generate data. The remaining indices still chose the Cont more frequently than the population model GLC2.
4) Among the four manipulated factors, the percentage of DIF items and between-group ability variance were the two factors that had a determinant impact on model performance and model selection. Increasing the percentage of DIF items combined with smaller between-group variance resulted in better parameter and classification recovery and improved correct model identification.
5) The proportion of the PLCs had a more significant effect when the data-generation model was discrete. In particular, for the GLC2, an even proportion improved model parameter recovery which in turn increased correct model identification.
6) Once the person-level sample size was fixed, the effect of group size was insignificant on model parameter recovery but moderate on model

selection. Large group size corresponded to a small number of groups. By reducing the number of groups, fit indices tended to favor complex discrete models with more GLCs, especially for the fit indices that incorporated sample size in their equations.

In the following sections, the implication of item bias and RMSE is discussed first, followed by the insights on the comparison of model performance. The research question regarding model selection in MMIRT using information-based fit indices is address in the last section.

5.1.1 Item Bias and RMSE.

Bias captures the degree to which a model deviates from the population values, while RMSE combines the information of bias and random variation to reflect overall model performance on parameter recovery. It was observed in the current study that the ranges of bias and RMSE were fairly close and even identical on the effects of manipulated factors.

Strong correlation between bias and RMSE. RMSE is the square root of mean squared error (MSE), which can be decomposed into squared bias and variance. The variance reflects sampling fluctuation. A large MSE can be a result of either bias in the estimation or just a result of variation in the sample.

Based on Equation 3.13, a positive bias corresponds to overestimation and negative value to underestimation. The average bias was found positive across the simulated conditions. Further analysis indicated that the percentage of RMSE explained by bias was over 80% in the continuous model and 50% to 75% in the discrete models. Thus, overall the estimation models led to a systematic overestimate

of the difference between item difficulty parameters across latent classes, especially for the Cont model. The next question that should be asked is why all the models tended, on average, to overestimate the difference? The answer probably depends on whether or not the item has DIF effect.

Item recovery on different item types. Rather than taking all items as a whole, the current study differentiates between the non-DIF and DIF items when assessing the large bias and RMSE found on parameter recovery. The descriptive statistics of bias and RMSE for the two item types were fully presented in Table 6a to 6c in Appendix A.

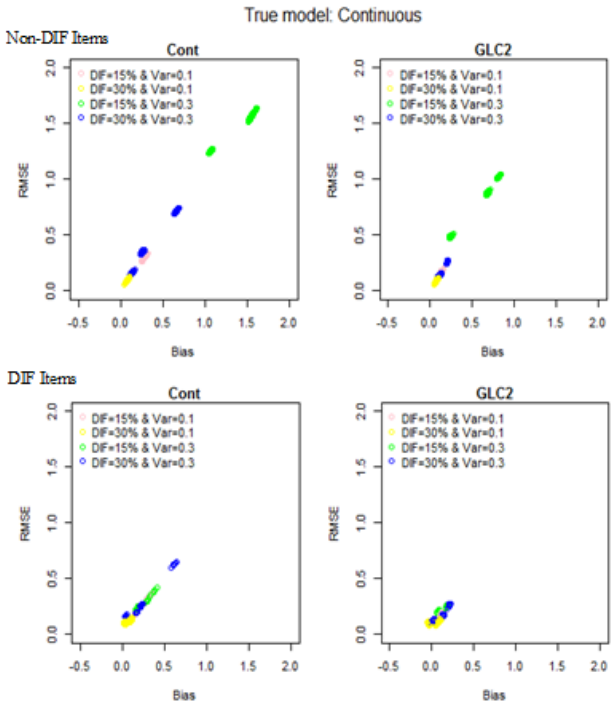

Figure 5.1. Scatterplots of item bias and RMSE on item types

A large discrepancy was frequently observed when a small percentage of DIF items were coupled with large between-group variance. To illustrate the relation between item bias and RMSE on different item types, two scatterplots between the

two criteria on the Cont and GLC2 were included in Figure 5.1. The scatterplots displayed almost a perfect correlation between bias and RMSE in non-DIF items across estimation models. Recalling the earlier results, the same conclusion can be reached in Figure 5.1 where the magnitude of bias and RMSE was the largest on non-DIF item for the Cont model under the condition with 15% of DIF items and between-group variance of 0.3. Such discrepancy was smallest on DIF items for the GLC2.

A plausible explanation for the consistent overestimation of difference on non-DIF items may be the insufficient information for separating the lower-level latent class. Put differently, a large between-group variance corresponds to small within-group variability, meaning individuals from the same group have a similar ability level. However, if those individuals show distinct response patterns on the same set of items, it is an indication of strong DIF effects. The problem is that such an effect is hard to achieve with only 15% of true DIF items. Thus, the item difficulty difference of the true non-DIF items is exaggerated in order to support the separation of latent classes.

This false-positive DIF effect can be a potential problem for the interpretation of MMIRT results. Note that in the empirical analysis, the items from MSA were also found to have significant difference across student-level latent classes. Differentiation on item functioning can be a threat to test reliability (Mislevy & Verhelst, 1990) and construct validity (Cho, 2007), and should be controlled in conventional tests. When a majority of items are identified by MMIRT models to have DIF effect, practitioners

should be cautious and consider that as an indicator for the difficulty of separating latent classes.

5.1.2 Comparison of Model Performance.

One surprising but worrisome finding in the simulation study is that, except for the GLC2 model, the Cont and GLC4 performed poorly on parameter and classification recovery, even when they were used to generate data.

A possible reason could be that the conditions simulated in the current study were not optimal for the performance of complex MMIRT models such as Cont and GLC4. As a complex modeling framework, most existing literature has only limited their discussion of multilevel mixture models on model introduction and empirical illustration (e.g., Asparouhov & Muthén, 2008; Palardy & Vermunt, 2010; Van Horn et al., 2008; Varriale & Vermunt, 2012; Vermunt, 2003, 2007, 2008b). A systematic evaluation using simulated data has been only conduced in one study so far. Cho (2007) examined the performance of discrete MMIRT models under practical DIF testing conditions. Three distinctive differences between Cho's study and this simulation study, however, prevented a comparison of her findings in discrete MMIRT models to the current study. First of all, in Cho's design, DIF items were not only introduced at the student level but also at the school level. Second, all items were generated to have DIF size ranging from 0.5 to 3 between student-level latent classes. Third, only two latent classes were considered at both the student level and the school level. Even with such a large difference in item difficulty across latent classes, the model can only well recover model parameters with 30% school-level DIF.

Compared to the Cho's study, the current study considered more realistic settings. This simulation design was based on a typical state assessment in which large DIF effect is expected to be eliminated for test reliability and validity reasons. Hence, only a percentage of the moderate DIF items across person-level latent classes are considered in the current simulation study. Apparently under such specifications, it is rather challenging to identify DIF items; further separate latent classes at the person level, and eventually recover random effects at the group level. For the Cont and GLC4, which are more complex models, the estimation can be even harder under the current simulation design.

The high non-convergence rate commonly seen among the MMIRT models also is an indication of the difficulty for model estimation under varying simulated conditions. However, one interesting finding is that the Cont model can converge most of time, which supports the early argument that a continuous latent distribution can well reproduce discrete latent distribution (Haertel, 1990; Markon & Krueger, 2006). Despite the potential violation of normality assumption, another appealing advantage of the continuous approach to approximating discrete distribution is that it can avoid the possible empty class as frequently encountered in discrete models.

5.1.3 Model Selection in MMIRT.

One main purpose of the current study was to find the information criterion or criteria that can successfully differentiate between the continuous and discrete MMIRT models which differ only with respect to the specification group-level random effects.

Markon and Krueger (2006) remarked previously that even for a non-normal latent distribution, a normal distribution might be preferred over a discrete distribution under certain conditions, since the loss of statistical information about the observed sample is less under the normal distribution. While the results from the simulation study indicated that if the comparison was made between the continuous model and the discrete model, the former one was generally preferred over the latter by the fit indices, regardless of the true distribution. On the other hand, if the distribution was severely discrete, although correct model selection increased in various fit indices, the selection rate was not as high as in the continuous model. From a different aspect, the findings above indicate how difficult it is to distinguish between the continuous and discrete MMIRT models when the actual distribution is moderately to mildly discrete. For practitioners who want to use information-based fit indices to infer the discreteness versus continuousness of higher-level random effects, this finding provides strong evidence against doing that. In particular, the current simulation only included the discrete distributions that are relatively symmetric, similar to what occurs with a normal distribution. Therefore, the symmetry assumption was not severely violated. A greater violation of the symmetry assumption might lead to more distinctive model performance between the two types of MMIRT models.

Information-based fit index. The simulation study was also cautioned to only rely on one fit index to make a selection decision. Three index groups have been identified, which are characterized by distinctive selection patterns across simulated conditions.

The AIC, AIC3, AICC are the three indices found to be most sensitive to discreteness. When the true distribution of random effects is discrete, such as the GLC2 and GLC4, these three indices tend to choose the optimal solution among the discrete models. If restricting the comparison to discrete models only, another interesting feature of these three indices is that they more frequently selected the true model than did the other indices. This finding contradicts to what is known about the AIC in traditional mixture models, where it always favors complex models (Lubke & Neale, 2006; McLachlan & Peel, 2000).

The second type of indices includes the CAIC, BIC, ssBIC, nCAIC and nBIC. These indices consistently prefer the continuous models over the discrete ones. Even when the population model was the GLC2, only through a large percentage of DIF items and small between-group variance can these indices choose the GLC2 over the Cont. The findings here support the argument that the CAIC and the BIC always reach the same selection with large sample size (Lubke & Neale, 2006; Markon & Krueger, 2006). Replacing total sample size with number of groups in the BIC, this change seems to result in no improvement on model selection in the current simulation. In contrast, the new version of ssBIC functioned better than the original index in capturing discreteness in the distribution.

It is not surprising that the performance of the modified versions of AICC and ssBIC depends on the sample size at a higher level. The nAICC, for example, functioned similarly with the BIC type indices when the number of groups was large. Once the group number became smaller sometimes even smaller than the number of parameters in the current simulation, the nAICC performance was close to the AIC

type indices. This finding indicates that the modified AICC should not be used given its unstable function under varying number of groups. The nssBIC, to the contrary, is less affected by the group number. Its model selection decision was similar with the AIC type indices under large group number. A smaller group number led to a higher preference of more complex discrete MMIRTs in the nssBIC than in the AIC.

Comparison of competing models. Hamaker et al. (2011) commented on the model fit in multilevel models that "... it is not the size of the information criterion that matters. Rather, it is the difference between information criteria for competing models that is of interest" (p. 233). The current study adopted a likelihood ratio approach to show the size of fit difference between two competing models with the smallest fit values.

The simulation results showed consistent patterns on the difference across the fit indices. When the comparison was between the continuous model and discrete model, the value of ΔIC was often larger than 9 if the Cont was chosen as the best fitting model. On the other hand, if the discrete model was preferred over the Cont, the value of ΔIC was fairly small. Restricting comparison to discrete models would always result in choosing the one with smallest number of GLCs, regardless of the population model. Both findings can be explained by at least two possible reasons. First, the model log-likelihood does not differ much across the estimation models; second, the number of parameters is larger in the discrete models than in the continuous model, when the same number of PLCs is specified at a lower-level. The way the information-based criteria are formulated leads to a consistent preference to the continuous models.

When discussing model fit indices in the context of mixture models, the existing literature focuses mainly on their performance in identifying correct number of latent class at a lower level. No study has looked at their function when the mixtures are at a higher level. The lack of systematic evaluation of fit indices from the previous research suggests that more studies are needed to address this problem.

5.2 Application of MMIRT models

The multilevel mixture models enable researchers to characterize group heterogeneity among higher-level units in terms of the latent attributes of their lower-level units. Previously, MLCA was proved to allow for the assessment of latent class typologies in contextual studies (Henry & Muthén, 2010). The MMIRT models incorporate traditional IRT models into MLCA, and extend their applications to a broader scenario where both latent class and continuous latent ability can be utilized to describe the variation across higher-level units.

The current simulation design is based on the findings of traditional mixture models and multilevel models. The influential factors examined previously do not provide optimal conditions under which the MMIRT models are able to well recover the population parameters. Of the four manipulated factors, three are characteristic of lower-level mixtures over which researchers have little control in a real setting. Fortunately, a large percentage of DIF items can help increase the identification of latent class membership and recover item parameters. What is more important, the negative impact of mixture characteristics, such as an uneven proportion of PLCs and large group heterogeneity on latent ability, can be counterbalanced by introducing more DIF items in the test. Hence, for the proper application of MMIRT models, a

well-designed measurement should be the most important requirement. If a test is not designed to differentiate potential latent groups, like the state assessment included in the empirical application, researchers should be cautious about the inference of results obtained from MMIRT models. For instance, the overestimated difficulty difference was pervasive in the non-DIF items, suggesting that a false-positive DIF effect would be expected on test items.

With regard to model comparison in MMIRT, the fact that the continuous model performed so poorly on model parameter and classification recovery in the simulation provides compelling evidence against selecting this model as the best solution. However, since population values are unknown in empirical data, discriminating between a continuous MMIRT model and a discrete one with large number of GLCs can be tough. In particular, classifying higher-level units in a discrete model may result in an empty class, which further increases the challenge to identify the discreteness at the group level. A direct comparison between the two approaches to MMIRT modeling may reach misleading conclusions. Therefore, before the application of either approach to MMIRT models, practitioners should have a sound theoretical foundation to support the choice of random effect distribution.

Which criteria should be used in distinguishing between MMIRT models can be an even tougher question. Because the decision of model selection is not purely a statistical issue, it also requires tremendous judgments about the research purposes and the nature of social reality (Weaklim, 2004). If the model comparison is to explore an unknown distribution with limited supportive evidence, the fit indices such

as the AIC, AIC3, AICC, and nssBIC can be a better indicator of discreteness in distribution. Meanwhile, if previous findings strongly support a continuous distribution, the BIC, CAIC, and ssBIC can successfully identify the true structure.

5.3 Limitations and Future Direction

The current study is the first to introduce continuous and discrete MMIRT models that differ only at the specifications of higher-level random effects. For this new modeling approach, existing literature rarely evaluated model performance under a variety of conditions using a simulation study.

As discussed earlier, due to the absence of a systematic evaluation of multilevel mixture models, the conditions manipulated in the current simulation might be inadequate to differentiate the model performance of the continuous and discrete MMIRT models. Especially for the complex discrete models, the high non-convergence rates commonly seen across simulated conditions did not support Heinen (1996) and Vermunt's (2001) comments that more discrete latent classes can approximate continuous model. Although the DIF effect significantly impacts parameter recovery and model selection, more research is required to understand how large the DIF effect size should be for a stable separation of lower-level latent classes. Moreover, the total sample size is sufficiently large under the current model specification. The effect of reduced sample size on model performance is unclear. In particular, what are the minimum sample sizes between and within higher-level units for stable model estimation? Both the Cho study and the current study only discussed two latent classes at the lower-level. It is interesting to assess how the identification

of higher-level random effects is affected by increasing number of lower-level latent classes.

Model selection is a complex process in MMIRT, especially in the discrete models. In the simulation study, the specification of a lower-level mixture model was the same as the data-generation model. Model misspecification only occurred at the group level. Therefore, the interaction of model misspecification between the lower and higher level is still of interest. This issue is relevant to the decision making process involved in MMIRT models. Henry and Muthén (2010) suggested ignoring hierarchical structure and deciding the number of lower-level latent classes first using a traditional mixture model. Additional group-level random effects are included afterwards. However, as observed in the empirical application with the same number of higher-level latent classes, the fit indices might change their preference on the number of lower-level latent classes. Apparently the specification of random effects at group-level can lead to a substantial change in the decision of model selection.

Given the substantial computing time required for model estimation in MMIRT, the current study constrained the number of sets of random starting values and total replications in order to complete the simulation study within a manageable time period. Those changes can potentially threaten the generalizability of findings in the current study. Muthén and Muthén (1998-2010) recommended using more sets of starting values, such as 100 sets of initial stage starting values and 10 for final stage optimization, for complex mixture models. The corresponding numbers are only 3 and 1 in the current study. On the other hand, the previous studies (e.g., Bauer & Curran, 2004; Lubke and Neale, 2006; Markon & Krueger, 2006), when using model

comparison to distinguish between the discrete and continuous latent variable models, usually generated hundreds of datasets in their simulation studies. The current study only used 50 replications, which is a relatively small number for this type of study. To better understand model function of MMIRT models, future study should consider including more replications in simulations and increasing the number of starting values at the initial and final stages in ML.

The current study only discussed the performance of MMIRT models without covariates. In fact, an increasing number of studies advocated including covariates in model estimation for mixture models. For instance, the identified higher-level latent class in empirical application can be characterized by possible background information. Such a feature has an even more profound impact in educational setting. The use of MMIRT model can be promising for practitioners to identify and explain why some teachers or schools are associated with students with similar strengths or weaknesses on a particular subject area or a designated skill.

Appendix A

Table 1
Generated item difficulty parameter and DIF items

		DIF Effect in Class 2				DIF Effect in Class 2	
	Class 1	15%	30%		Class 1	15%	30%
Item 1	-1.448	0	0	Item 21	-0.042	0	0
Item 2	-1.259	0	0	Item 22	-0.012	0	1.00
Item 3	-1.182	0	0	Item 23	0.072	0	0
Item 4	-1.150	0	0	Item 24	0.205	0	0
Item 5	-1.140	0	0	Item 25	0.330	1.00	1.00
Item 6	-1.095	0	0	Item 26	0.337	0	0
Item 7	-1.034	0	0	Item 27	0.379	0	1.00
Item 8	-1.003	0	1.00	Item 28	0.612	0	0
Item 9	-0.981	0	0	Item 29	0.654	0	0
Item 10	-0.819	1.00	1.00	Item 30	0.723	1.00	1.00
Item 11	-0.682	0	0	Item 31	0.723	0	0
Item 12	-0.664	0	1.00	Item 32	0.767	0	1.00
Item 13	-0.642	0	0	Item 33	0.810	0	0
Item 14	-0.633	0	0	Item 34	0.900	0	0
Item 15	-0.467	1.00	1.00	Item 35	0.920	1.00	1.00
Item 16	-0.459	0	0	Item 36	0.978	0	0
Item 17	-0.286	0	1.00	Item 37	1.229	0	0
Item 18	-0.250	0	0	Item 38	1.397	0	0
Item 19	-0.215	0	0	Item 39	1.408	0	0
Item 20	-0.050	1.00	1.00	Item 40	1.442	0	0

Table 2a

Descriptive statistics of item parameter bias (True model: Continuous)

				Estimation Model							
				Cont		GLC2		GLC3		GLC4	
DIF	Size	Prop	Var	M	SD	M	SD	M	SD	M	SD
15%	25	50/50	0.1	0.13	0.01	0.08	0.01	0.07	0.01	0.07	0.01
			0.3	1.38	0.43	0.22	0.06	0.14	0.05	0.68	0.22
		30/70	0.1	0.14	0.01	0.12	0.01	0.12	0.01	0.11	0.01
			0.3	0.93	0.32	0.60	0.22	0.75	0.26	1.12	0.41
	150	50/50	0.1	0.30	0.02	0.13	0.01	0.15	0.01	0.16	0.01
			0.3	1.39	0.45	0.71	0.25	0.89	0.31	1.08	0.38
		30/70	0.1	0.27	0.02	0.13	0.01	0.17	0.01	0.20	0.01
			0.3	1.35	0.44	0.60	0.18	1.19	0.40	1.22	0.42
30%	25	50/50	0.1	0.10	0.01	0.07	0.01	0.08	0.01	0.08	0.01
			0.3	0.15	0.01	0.13	0.01	0.14	0.01	0.14	0.01
		30/70	0.1	0.09	0.01	0.06	0.01	0.05	0.02	0.05	0.02
			0.3	0.10	0.04	0.04	0.05	0.05	0.04	0.06	0.05
	150	50/50	0.1	0.05	0.02	0.04	0.05	0.05	0.06	0.05	0.07
			0.3	0.24	0.02	0.07	0.04	0.11	0.04	0.18	0.03
		30/70	0.1	0.09	0.01	0.08	0.01	0.06	0.01	0.06	0.01
			0.3	0.64	0.03	0.20	0.01	0.51	0.02	0.55	0.02

Table 2b

Descriptive statistics of item parameter bias (True model: GLC2)

				Estimation Model							
				Cont		GLC2		GLC3		GLC4	
DIF	Size	Prop	Var	M	SD	M	SD	M	SD	M	SD
15%	25	50/50	0.1	0.08	0.01	0.06	0.02	0.06	0.02	0.06	0.01
			0.3	0.45	0.02	0.22	0.02	0.30	0.02	0.31	0.02
		30/70	0.1	0.11	0.01	0.10	0.01	0.10	0.01	0.09	0.01
			0.3	0.89	0.27	0.57	0.18	0.72	0.23	0.95	0.32
	150	50/50	0.1	0.06	0.02	0.07	0.06	0.08	0.07	0.08	0.07
			0.3	0.37	0.04	0.13	0.03	0.16	0.03	0.17	0.03
		30/70	0.1	0.15	0.01	0.08	0.02	0.08	0.01	0.09	0.02
			0.3	1.21	0.40	0.57	0.17	1.02	0.33	0.96	0.31
30%	25	50/50	0.1	0.04	0.05	0.04	0.06	0.04	0.06	0.04	0.06
			0.3	0.04	0.07	0.06	0.13	0.07	0.14	0.06	0.13
		30/70	0.1	0.12	0.01	0.07	0.01	0.06	0.01	0.06	0.02
			0.3	0.39	0.02	0.10	0.01	0.08	0.02	0.07	0.02
	150	50/50	0.1	0.28	0.02	0.17	0.01	0.20	0.01	0.20	0.01
			0.3	0.18	0.02	0.03	0.02	0.04	0.01	0.04	0.01
		30/70	0.1	0.05	0.04	0.05	0.04	0.05	0.06	0.06	0.07
			0.3	0.60	0.02	0.11	0.01	0.53	0.01	0.58	0.02

Table 2c

Descriptive statistics of item parameter bias (True model: GLC4)

				Estimation Model							
				Cont		GLC2		GLC3		GLC4	
DIF	Size	Prop	Var	M	SD	M	SD	M	SD	M	SD
15%	25	50/50	0.1	0.11	0.01	0.08	0.01	0.08	0.01	0.08	0.01
			0.3	0.48	0.01	0.24	0.01	0.12	0.02	0.21	0.02
		30/70	0.1	0.19	0.02	0.14	0.01	0.13	0.01	0.13	0.01
			0.3	1.23	0.37	1.10	0.35	1.08	0.35	1.12	0.36
	150	50/50	0.1	0.19	0.02	0.09	0.01	0.09	0.01	0.10	0.01
			0.3	1.22	0.25	0.90	0.11	1.01	0.15	1.13	0.20
		30/70	0.1	0.10	0.01	0.08	0.06	0.09	0.07	0.09	0.07
			0.3	1.28	0.53	1.14	0.51	1.23	0.54	1.28	0.55
30%	25	50/50	0.1	0.06	0.01	0.05	0.02	0.05	0.02	0.05	0.02
			0.3	0.17	0.01	0.15	0.01	0.15	0.01	0.14	0.01
		30/70	0.1	0.07	0.01	0.06	0.02	0.06	0.02	0.05	0.02
			0.3	0.38	0.01	0.23	0.02	0.31	0.02	0.29	0.02
	150	50/50	0.1	0.12	0.01	0.07	0.01	0.07	0.01	0.08	0.01
			0.3	0.11	0.21	0.09	0.22	0.09	0.23	0.10	0.24
		30/70	0.1	0.06	0.01	0.05	0.05	0.05	0.05	0.05	0.05
			0.3	1.35	0.45	0.63	0.13	0.98	0.29	1.18	0.39

Table 3a

Descriptive statistics of item parameter RMSE (True model: Continuous)

				Estimation Model							
				Cont		GLC2		GLC3		GLC4	
DIF	Size	Prop	Var	M	SD	M	SD	M	SD	M	SD
15%	25	50/50	0.1	0.16	0.02	0.10	0.01	0.09	0.01	0.10	0.01
			0.3	1.40	0.43	0.44	0.13	0.34	0.09	0.91	0.30
		30/70	0.1	0.17	0.02	0.16	0.02	0.16	0.01	0.15	0.01
			0.3	1.10	0.37	0.79	0.25	0.94	0.30	1.17	0.41
	150	50/50	0.1	0.32	0.02	0.17	0.02	0.20	0.01	0.22	0.02
			0.3	1.41	0.46	0.90	0.31	1.04	0.36	1.16	0.40
		30/70	0.1	0.29	0.02	0.17	0.01	0.20	0.02	0.23	0.02
			0.3	1.35	0.44	0.78	0.23	1.21	0.40	1.23	0.41
30%	25	50/50	0.1	0.11	0.01	0.09	0.01	0.10	0.01	0.10	0.01
			0.3	0.18	0.01	0.16	0.01	0.17	0.01	0.17	0.01
		30/70	0.1	0.11	0.02	0.09	0.01	0.08	0.01	0.08	0.01
			0.3	0.15	0.01	0.09	0.01	0.10	0.01	0.12	0.01
	150	50/50	0.1	0.08	0.01	0.09	0.01	0.10	0.01	0.11	0.01
			0.3	0.32	0.04	0.13	0.01	0.20	0.01	0.26	0.01
		30/70	0.1	0.11	0.02	0.10	0.02	0.09	0.01	0.08	0.01
			0.3	0.69	0.04	0.25	0.01	0.52	0.02	0.56	0.02

Table 3b

Descriptive statistics of item parameter RMSE (True model: GLC2)

				Estimation Model							
				Cont		GLC2		GLC3		GLC4	
DIF	Size	Prop	Var	M	SD	M	SD	M	SD	M	SD
15%	25	50/50	0.1	0.10	0.01	0.09	0.01	0.09	0.01	0.08	0.01
			0.3	0.46	0.02	0.25	0.02	0.34	0.02	0.35	0.02
		30/70	0.1	0.14	0.02	0.14	0.02	0.13	0.01	0.13	0.01
			0.3	1.04	0.30	0.72	0.20	0.89	0.27	1.04	0.34
	150	50/50	0.1	0.08	0.01	0.11	0.01	0.12	0.01	0.12	0.01
			0.3	0.38	0.04	0.15	0.03	0.17	0.03	0.19	0.03
		30/70	0.1	0.18	0.02	0.12	0.01	0.12	0.01	0.12	0.01
			0.3	1.21	0.40	0.73	0.21	1.07	0.34	1.04	0.33
30%	25	50/50	0.1	0.08	0.01	0.09	0.01	0.09	0.01	0.09	0.01
			0.3	0.10	0.01	0.16	0.01	0.16	0.01	0.16	0.01
		30/70	0.1	0.14	0.01	0.09	0.01	0.08	0.01	0.09	0.01
			0.3	0.41	0.02	0.18	0.01	0.15	0.01	0.14	0.01
	150	50/50	0.1	0.29	0.02	0.18	0.01	0.21	0.01	0.21	0.01
			0.3	0.20	0.02	0.06	0.01	0.06	0.01	0.06	0.01
		30/70	0.1	0.09	0.01	0.09	0.01	0.10	0.01	0.12	0.01
			0.3	0.61	0.02	0.16	0.01	0.55	0.02	0.60	0.02

Table 3c

Descriptive statistics of item parameter RMSE (True model: GLC4)

				Estimation Model							
				Cont		GLC3		GLC4		GLC5	
DIF	Size	Prop	Var	M	SD	M	SD	M	SD	M	SD
15%	25	50/50	0.1	0.13	0.02	0.11	0.01	0.11	0.01	0.11	0.01
			0.3	0.53	0.02	0.35	0.01	0.25	0.03	0.38	0.05
		30/70	0.1	0.22	0.02	0.17	0.02	0.15	0.02	0.16	0.02
			0.3	1.23	0.37	1.13	0.36	1.11	0.35	1.13	0.36
	150	50/50	0.1	0.20	0.02	0.11	0.01	0.12	0.01	0.13	0.01
			0.3	1.25	0.25	0.94	0.12	1.05	0.15	1.17	0.21
		30/70	0.1	0.13	0.01	0.13	0.01	0.15	0.01	0.15	0.01
			0.3	1.32	0.50	1.17	0.44	1.26	0.48	1.30	0.50
30%	25	50/50	0.1	0.08	0.01	0.08	0.01	0.08	0.01	0.08	0.01
			0.3	0.18	0.01	0.17	0.01	0.17	0.01	0.16	0.01
		30/70	0.1	0.10	0.01	0.09	0.01	0.09	0.01	0.09	0.01
			0.3	0.40	0.01	0.30	0.01	0.35	0.01	0.34	0.01
	150	50/50	0.1	0.13	0.01	0.09	0.01	0.09	0.01	0.10	0.01
			0.3	0.32	0.01	0.27	0.01	0.27	0.01	0.28	0.01
		30/70	0.1	0.08	0.01	0.09	0.01	0.09	0.01	0.09	0.01
			0.3	1.38	0.46	0.76	0.19	1.08	0.34	1.24	0.42

Table 4a

Descriptive statistics of person-level latent class classification recovery (True model: Continuous)

				Estimation Model							
				Cont*		GLC2		GLC3		GLC4	
DIF	Size	Prop	Var	M	SD	M	SD	M	SD	M	SD
15%	25	50/50	0.1	0.52	0.01	0.51	0.01	0.51	0.01	0.51	0.01
			0.3	0.02	0.07	0.45	0.17	0.48	0.12	0.25	0.25
		30/70	0.1	0.51	0.02	0.50	0.02	0.50	0.02	0.50	0.02
			0.3	0.21	0.21	0.28	0.22	0.25	0.22	0.09	0.14
	150	50/50	0.1	0.51	0.02	0.51	0.02	0.52	0.01	0.52	0.01
			0.3	0.08	0.08	0.24	0.23	0.19	0.21	0.12	0.16
		30/70	0.1	0.51	0.02	0.51	0.02	0.51	0.02	0.51	0.02
			0.3	-0.03	0.02	0.25	0.25	0.01	0.09	-0.01	0.01
30%	25	50/50	0.1	0.65	0.01	0.64	0.01	0.64	0.01	0.64	0.01
			0.3	0.64	0.01	0.64	0.01	0.64	0.01	0.64	0.01
		30/70	0.1	0.63	0.01	0.62	0.01	0.62	0.01	0.62	0.01
			0.3	0.62	0.01	0.62	0.01	0.62	0.01	0.62	0.01
	150	50/50	0.1	0.66	0.01	0.65	0.01	0.65	0.01	0.65	0.01
			0.3	0.62	0.09	0.64	0.01	0.64	0.03	0.63	0.04
		30/70	0.1	0.64	0.01	0.62	0.01	0.64	0.01	0.64	0.01
			0.3	0.55	0.12	0.62	0.02	0.59	0.03	0.59	0.03

Table 4b

Descriptive statistics of person-level latent class classification recovery (True model: GLC2)

				Estimation Model							
				Cont		GLC2*		GLC3		GLC4	
DIF	Size	Prop	Var	M	SD	M	SD	M	SD	M	SD
15%	25	50/50	0.1	0.64	0.01	0.63	0.06	0.66	0.02	0.65	0.02
			0.3	0.57	0.03	0.61	0.04	0.60	0.04	0.60	0.04
		30/70	0.1	0.49	0.02	0.49	0.02	0.49	0.02	0.49	0.02
			0.3	0.17	0.23	0.26	0.24	0.21	0.24	0.11	0.20
	150	50/50	0.1	0.67	0.01	0.68	0.01	0.68	0.01	0.68	0.01
			0.3	0.64	0.02	0.68	0.01	0.68	0.01	0.68	0.01
		30/70	0.1	0.50	0.02	0.51	0.01	0.51	0.01	0.51	0.01
			0.3	0.02	0.01	0.26	0.25	0.08	0.16	0.11	0.19
30%	25	50/50	0.1	0.73	0.01	0.74	0.01	0.74	0.01	0.73	0.01
			0.3	0.72	0.01	0.73	0.01	0.73	0.01	0.73	0.01
		30/70	0.1	0.62	0.01	0.61	0.01	0.61	0.01	0.61	0.01
			0.3	0.59	0.02	0.61	0.02	0.61	0.02	0.61	0.02
	150	50/50	0.1	0.73	0.01	0.75	0.01	0.74	0.01	0.74	0.01
			0.3	0.72	0.01	0.75	0.01	0.75	0.01	0.74	0.01
		30/70	0.1	0.63	0.01	0.63	0.01	0.63	0.01	0.63	0.01
			0.3	0.57	0.03	0.62	0.01	0.58	0.03	0.57	0.03

Table 4c

Descriptive statistics of person-level latent class classification recovery (True model: GLC4)

				Estimation Model							
				Cont		GLC3		GLC4*		GLC5	
DIF	Size	Prop	Var	M	SD	M	SD	M	SD	M	SD
15%	25	50/50	0.1	0.56	0.01	0.53	0.03	0.52	0.02	0.52	0.02
			0.3	0.50	0.10	0.50	0.07	0.50	0.07	0.49	0.12
		30/70	0.1	0.51	0.01	0.51	0.02	0.51	0.02	0.51	0.02
			0.3	-0.04	0.02	-0.02	0.11	-0.01	0.13	-0.03	0.08
	150	50/50	0.1	0.59	0.01	0.58	0.01	0.59	0.01	0.59	0.01
			0.3	0.22	0.14	0.37	0.15	0.32	0.15	0.27	0.16
		30/70	0.1	0.53	0.02	0.53	0.02	0.53	0.02	0.53	0.02
			0.3	0.18	0.07	0.16	0.04	0.18	0.04	0.18	0.04
30%	25	50/50	0.1	0.68	0.01	0.67	0.01	0.67	0.01	0.67	0.01
			0.3	0.66	0.01	0.66	0.01	0.66	0.01	0.66	0.01
		30/70	0.1	0.65	0.01	0.63	0.02	0.65	0.02	0.65	0.02
			0.3	0.61	0.02	0.61	0.02	0.61	0.02	0.61	0.02
	150	50/50	0.1	0.69	0.01	0.69	0.01	0.69	0.01	0.69	0.01
			0.3	0.67	0.03	0.67	0.02	0.68	0.02	0.68	0.02
		30/70	0.1	0.66	0.01	0.66	0.01	0.66	0.01	0.66	0.01
			0.3	0.11	0.16	0.47	0.22	0.28	0.25	0.18	0.21

Table 5a

Descriptive statistics of correlations between the true and estimated proportion of person membership within groups (True model: Continuous)

				Estimation Model							
				Cont*		GLC2		GLC3		GLC4	
DIF	Size	Prop	Var	M	SD	M	SD	M	SD	M	SD
15%	25	50/50	0.1	0.76	0.03	0.73	0.03	0.73	0.03	0.73	0.03
			0.3	0.03	0.10	0.61	0.22	0.65	0.16	0.35	0.34
		30/70	0.1	0.75	0.03	0.72	0.03	0.72	0.03	0.72	0.03
			0.3	0.33	0.29	0.43	0.26	0.37	0.29	0.16	0.18
	150	50/50	0.1	0.80	0.05	0.82	0.05	0.82	0.05	0.82	0.05
			0.3	0.22	0.06	0.42	0.26	0.36	0.27	0.26	0.20
		30/70	0.1	0.85	0.04	0.83	0.04	0.84	0.04	0.84	0.04
			0.3	0.07	0.02	0.46	0.36	0.04	0.14	0.02	0.02
30%	25	50/50	0.1	0.86	0.02	0.83	0.02	0.84	0.02	0.84	0.02
			0.3	0.82	0.03	0.81	0.02	0.81	0.03	0.81	0.03
		30/70	0.1	0.84	0.02	0.82	0.02	0.82	0.02	0.82	0.02
			0.3	0.77	0.05	0.76	0.03	0.76	0.03	0.77	0.04
	150	50/50	0.1	0.95	0.01	0.93	0.01	0.94	0.01	0.95	0.01
			0.3	0.80	0.13	0.87	0.06	0.87	0.10	0.83	0.11
		30/70	0.1	0.92	0.02	0.90	0.02	0.91	0.02	0.92	0.02
			0.3	0.67	0.13	0.83	0.06	0.74	0.06	0.73	0.05

Table 5b

Descriptive statistics of correlations between the true and estimated proportion of person membership within groups (True model: GLC2)

				Estimation Model							
				Cont		GLC2*		GLC3		GLC4	
DIF	Size	Prop	Var	M	SD	M	SD	M	SD	M	SD
15%	25	50/50	0.1	0.93	0.01	0.92	0.03	0.93	0.02	0.93	0.02
			0.3	0.82	0.04	0.87	0.04	0.86	0.05	0.86	0.05
		30/70	0.1	0.65	0.04	0.64	0.03	0.64	0.04	0.64	0.04
			0.3	0.23	0.29	0.33	0.29	0.28	0.29	0.15	0.24
	150	50/50	0.1	0.99	0.00	1.00	0.00	1.00	0.00	1.00	0.00
			0.3	0.95	0.02	0.99	0.01	0.99	0.01	0.99	0.01
		30/70	0.1	0.80	0.06	0.84	0.07	0.83	0.07	0.84	0.07
			0.3	0.11	0.02	0.44	0.35	0.21	0.24	0.24	0.28
30%	25	50/50	0.1	0.96	0.00	0.96	0.01	0.96	0.01	0.96	0.01
			0.3	0.95	0.01	0.96	0.01	0.96	0.01	0.96	0.01
		30/70	0.1	0.77	0.02	0.75	0.02	0.75	0.02	0.75	0.02
			0.3	0.67	0.05	0.72	0.03	0.72	0.03	0.72	0.03
	150	50/50	0.1	0.98	0.00	1.00	0.00	0.99	0.00	0.99	0.00
			0.3	0.97	0.01	1.00	0.00	0.99	0.00	0.99	0.00
		30/70	0.1	0.92	0.02	0.91	0.03	0.94	0.02	0.93	0.02
			0.3	0.68	0.06	0.83	0.05	0.70	0.07	0.68	0.06

Table 5c

Descriptive statistics of correlations between the true and estimated proportion of person membership within groups (True model: GLC4)

				Estimation Model							
				Cont		GLC3		GLC4*		GLC5	
DIF	Size	Prop	Var	M	SD	M	SD	M	SD	M	SD
15%	25	50/50	0.1	0.85	0.02	0.82	0.03	0.81	0.02	0.81	0.02
			0.3	0.72	0.15	0.75	0.11	0.77	0.11	0.74	0.18
		30/70	0.1	0.76	0.02	0.75	0.03	0.75	0.03	0.75	0.03
			0.3	0.12	0.03	0.15	0.12	0.16	0.15	0.13	0.09
	150	50/50	0.1	0.95	0.02	0.94	0.02	0.95	0.02	0.95	0.02
			0.3	0.37	0.20	0.58	0.21	0.53	0.21	0.44	0.22
		30/70	0.1	0.92	0.02	0.93	0.02	0.93	0.02	0.94	0.02
			0.3	0.46	0.09	0.44	0.09	0.48	0.10	0.47	0.10
30%	25	50/50	0.1	0.91	0.01	0.91	0.01	0.90	0.01	0.90	0.01
			0.3	0.88	0.02	0.87	0.02	0.88	0.02	0.88	0.02
		30/70	0.1	0.87	0.02	0.86	0.02	0.87	0.02	0.87	0.02
			0.3	0.77	0.03	0.79	0.04	0.79	0.03	0.79	0.03
	150	50/50	0.1	0.97	0.01	0.97	0.01	0.98	0.01	0.98	0.01
			0.3	0.95	0.03	0.96	0.01	0.97	0.02	0.97	0.02
		30/70	0.1	0.96	0.01	0.97	0.01	0.97	0.01	0.97	0.01
			0.3	0.17	0.19	0.63	0.28	0.38	0.32	0.25	0.26

Table 6a

Descriptive statistics of item parameter bias and RMSE on item types (True mode: Continuous)

DIF	Size	Prop	Var	Non-DIF Items				DIF Items			
				Cont	GLC3	GLC4	GLC5	Cont	GLC3	GLC4	GLC5
Bias											
15%	25	50/50	0.1	0.13	0.08	0.07	0.07	0.15	0.08	0.06	0.05
			0.3	1.56	0.24	0.16	0.78	0.38	0.07	0.04	0.16
		30/70	0.1	0.14	0.12	0.12	0.11	0.15	0.13	0.11	0.10
			0.3	1.07	0.69	0.86	1.29	0.18	0.07	0.14	0.17
	150	50/50	0.1	0.30	0.13	0.15	0.16	0.34	0.15	0.16	0.17
			0.3	1.58	0.82	1.01	1.23	0.33	0.11	0.16	0.18
		30/70	0.1	0.26	0.13	0.17	0.20	0.29	0.11	0.17	0.21
			0.3	1.53	0.67	1.35	1.40	0.31	0.17	0.26	0.24
30%	25	50/50	0.1	0.09	0.07	0.08	0.08	0.11	0.07	0.08	0.08
			0.3	0.15	0.13	0.14	0.14	0.17	0.14	0.15	0.15
		30/70	0.1	0.08	0.07	0.06	0.06	0.10	0.05	0.03	0.03
			0.3	0.12	0.07	0.07	0.09	0.04	-0.03	-0.02	-0.02
	150	50/50	0.1	0.06	0.08	0.09	0.09	0.03	-0.04	-0.05	-0.05
			0.3	0.25	0.10	0.13	0.20	0.22	0.02	0.05	0.14
		30/70	0.1	0.08	0.07	0.07	0.06	0.09	0.09	0.06	0.06
			0.3	0.66	0.20	0.50	0.54	0.61	0.21	0.52	0.56
RMSE											
15%	25	50/50	0.1	0.16	0.10	0.09	0.09	0.19	0.12	0.11	0.11
			0.3	1.58	0.49	0.37	1.03	0.39	0.13	0.13	0.21
		30/70	0.1	0.17	0.15	0.15	0.15	0.19	0.18	0.17	0.17
			0.3	1.25	0.89	1.07	1.34	0.24	0.20	0.24	0.20
	150	50/50	0.1	0.31	0.16	0.19	0.21	0.36	0.19	0.22	0.24
			0.3	1.60	1.03	1.19	1.33	0.34	0.16	0.19	0.21
		30/70	0.1	0.28	0.17	0.20	0.23	0.31	0.18	0.22	0.25
			0.3	1.54	0.88	1.37	1.40	0.32	0.25	0.28	0.25
30%	25	50/50	0.1	0.11	0.09	0.10	0.10	0.13	0.10	0.11	0.12
			0.3	0.17	0.15	0.17	0.17	0.19	0.17	0.19	0.19
		30/70	0.1	0.10	0.08	0.08	0.08	0.13	0.10	0.09	0.09
			0.3	0.15	0.08	0.10	0.12	0.15	0.09	0.11	0.13
	150	50/50	0.1	0.08	0.09	0.11	0.11	0.09	0.09	0.10	0.10
			0.3	0.34	0.13	0.21	0.26	0.26	0.13	0.20	0.26
		30/70	0.1	0.10	0.09	0.08	0.08	0.13	0.12	0.10	0.09
			0.3	0.71	0.25	0.52	0.56	0.62	0.26	0.53	0.57

Table 6b

Descriptive statistics of item parameter bias and RMSE on item types (True mode: GLC2)

DIF	Size	Prop	Var	Non-DIF Items				DIF Items			
				Cont	GLC3	GLC4	GLC5	Cont	GLC3	GLC4	GLC5
Bias											
15%	25	50/50	0.1	0.08	0.07	0.07	0.06	0.06	0.02	0.02	0.03
			0.3	0.44	0.22	0.30	0.30	0.48	0.26	0.34	0.35
		30/70	0.1	0.11	0.10	0.10	0.10	0.10	0.10	0.09	0.08
			0.3	1.01	0.65	0.81	1.08	0.26	0.15	0.18	0.20
	150	50/50	0.1	0.06	0.10	0.11	0.11	0.01	-0.07	-0.09	-0.08
			0.3	0.35	0.12	0.15	0.16	0.45	0.20	0.23	0.25
		30/70	0.1	0.15	0.09	0.09	0.09	0.16	0.05	0.06	0.05
			0.3	1.38	0.64	1.16	1.09	0.27	0.17	0.23	0.23
30%	25	50/50	0.1	0.07	0.08	0.08	0.08	-0.03	-0.05	-0.05	-0.05
			0.3	0.08	0.15	0.15	0.15	-0.06	-0.14	-0.14	-0.14
		30/70	0.1	0.12	0.07	0.06	0.06	0.12	0.05	0.04	0.03
			0.3	0.39	0.11	0.09	0.08	0.39	0.09	0.05	0.05
	150	50/50	0.1	0.28	0.17	0.20	0.20	0.30	0.19	0.22	0.22
			0.3	0.17	0.05	0.05	0.05	0.21	0.00	0.02	0.03
		30/70	0.1	0.08	0.07	0.09	0.11	0.00	0.00	-0.03	-0.05
			0.3	0.60	0.12	0.53	0.58	0.59	0.11	0.52	0.57
RMSE											
15%	25	50/50	0.1	0.10	0.09	0.09	0.08	0.11	0.09	0.10	0.09
			0.3	0.46	0.24	0.33	0.34	0.50	0.29	0.37	0.38
		30/70	0.1	0.14	0.13	0.12	0.12	0.17	0.16	0.15	0.15
			0.3	1.16	0.80	1.00	1.18	0.33	0.26	0.25	0.25
	150	50/50	0.1	0.08	0.11	0.13	0.12	0.09	0.10	0.11	0.11
			0.3	0.36	0.13	0.16	0.17	0.46	0.21	0.25	0.26
		30/70	0.1	0.18	0.12	0.12	0.12	0.21	0.13	0.13	0.14
			0.3	1.38	0.82	1.21	1.18	0.28	0.25	0.26	0.27
30%	25	50/50	0.1	0.08	0.10	0.10	0.10	0.07	0.08	0.08	0.08
			0.3	0.10	0.16	0.17	0.16	0.10	0.15	0.16	0.15
		30/70	0.1	0.14	0.09	0.08	0.08	0.15	0.09	0.09	0.09
			0.3	0.41	0.18	0.15	0.15	0.41	0.18	0.15	0.14
	150	50/50	0.1	0.28	0.18	0.21	0.21	0.31	0.20	0.23	0.23
			0.3	0.18	0.06	0.06	0.06	0.22	0.06	0.06	0.07
		30/70	0.1	0.09	0.09	0.11	0.13	0.08	0.08	0.09	0.11
			0.3	0.62	0.16	0.55	0.60	0.60	0.16	0.54	0.59

Table 6c

Descriptive statistics of item parameter bias and RMSE on item types (True mode: GLC4)

DIF	Size	Prop	Var	Non-DIF Items				DIF Items			
				Cont	GLC3	GLC4	GLC5	Cont	GLC3	GLC4	GLC5
Bias											
15%	25	50/50	0.1	0.11	0.08	0.08	0.08	0.14	0.09	0.08	0.09
			0.3	0.47	0.24	0.13	0.22	0.49	0.22	0.09	0.16
		30/70	0.1	0.19	0.13	0.12	0.12	0.23	0.16	0.15	0.15
			0.3	1.38	1.25	1.22	1.27	0.36	0.28	0.27	0.27
	150	50/50	0.1	0.18	0.09	0.09	0.10	0.22	0.09	0.10	0.12
			0.3	1.32	0.94	1.07	1.21	0.65	0.64	0.67	0.66
		30/70	0.1	0.10	0.10	0.12	0.12	0.08	-0.05	-0.08	-0.08
			0.3	1.50	1.35	1.46	1.50	0.05	-0.07	-0.03	0.00
30%	25	50/50	0.1	0.06	0.06	0.07	0.07	0.06	0.03	0.02	0.02
			0.3	0.16	0.14	0.14	0.13	0.18	0.16	0.16	0.15
		30/70	0.1	0.08	0.07	0.07	0.07	0.07	0.04	0.03	0.03
			0.3	0.38	0.24	0.32	0.30	0.37	0.21	0.30	0.28
	150	50/50	0.1	0.11	0.07	0.07	0.08	0.13	0.06	0.07	0.08
			0.3	0.25	0.23	0.24	0.25	-0.22	-0.24	-0.25	-0.26
		30/70	0.1	0.06	0.08	0.08	0.08	0.06	-0.03	-0.03	-0.03
			0.3	1.64	0.72	1.17	1.43	0.67	0.44	0.54	0.59
RMSE											
15%	25	50/50	0.1	0.13	0.10	0.10	0.10	0.17	0.13	0.13	0.13
			0.3	0.53	0.35	0.26	0.40	0.51	0.34	0.20	0.25
		30/70	0.1	0.21	0.16	0.15	0.15	0.25	0.20	0.19	0.19
			0.3	1.39	1.28	1.26	1.28	0.37	0.29	0.29	0.28
	150	50/50	0.1	0.20	0.11	0.12	0.13	0.24	0.13	0.14	0.16
			0.3	1.36	0.99	1.11	1.25	0.66	0.67	0.69	0.68
		30/70	0.1	0.13	0.13	0.15	0.15	0.16	0.12	0.14	0.14
			0.3	1.52	1.36	1.46	1.51	0.14	0.14	0.13	0.13
30%	25	50/50	0.1	0.08	0.08	0.08	0.08	0.10	0.09	0.09	0.09
			0.3	0.18	0.17	0.16	0.15	0.20	0.19	0.18	0.17
		30/70	0.1	0.10	0.08	0.09	0.08	0.11	0.09	0.10	0.09
			0.3	0.40	0.30	0.36	0.34	0.39	0.29	0.34	0.33
	150	50/50	0.1	0.13	0.09	0.09	0.09	0.14	0.10	0.10	0.11
			0.3	0.32	0.27	0.27	0.28	0.34	0.28	0.28	0.29
		30/70	0.1	0.08	0.10	0.10	0.10	0.10	0.08	0.08	0.08
			0.3	1.68	0.88	1.30	1.51	0.68	0.48	0.56	0.61

Appendix B

Table 7a

Non-convergence frequency in simulated conditions (True model: Continuous)

				Estimation Model				
DIF	Size	Prop	Var	Cont	GLC2	GLC3	GLC4	Total
15%	25	50/50	0.1	0	0	0	0	0
			0.3	1	0	0	0	1
		30/70	0.1	0	0	11	12	20
			0.3	0	0	10	0	10
	150	50/50	0.1	0	0	0	0	0
			0.3	0	4	4	2	8
		30/70	0.1	0	0	0	0	0
			0.3	0	1	1	0	2
30%	25	50/50	0.1	0	0	1	1	2
			0.3	0	0	0	2	2
		30/70	0.1	0	0	0	1	1
			0.3	0	0	0	2	2
	150	50/50	0.1	0	0	0	0	0
			0.3	0	0	1	1	2
		30/70	0.1	0	0	0	0	0
			0.3	0	0	0	0	0

Table 7b

Non-convergence frequency in simulated conditions (True model: GLC2)

				Estimation Model				
DIF	Size	Prop	Var	Cont	GLC2	GLC3	GLC4	Total
15%	25	50/50	0.1	0	0	1	25	26
			0.3	0	0	3	6	7
		30/70	0.1	0	0	8	12	18
			0.3	0	0	7	5	11
	150	50/50	0.1	0	0	0	0	0
			0.3	0	0	0	0	0
		30/70	0.1	1	0	0	0	1
			0.3	0	6	4	2	10
30%	25	50/50	0.1	0	0	1	5	6
			0.3	0	0	0	0	0
		30/70	0.1	0	0	0	0	0
			0.3	0	0	0	1	1
	150	50/50	0.1	0	0	0	0	0
			0.3	0	0	0	0	0
		30/70	0.1	0	0	0	0	0
			0.3	28	0	8	15	32

Table 7c

Non-convergence frequency in simulated conditions (True model: GLC4)

				Estimation Model				
DIF	Size	Prop	Var	Cont	GLC3	GLC4	GLC5	Total
15%	25	50/50	0.1	0	0	4	5	7
			0.3	0	0	1	4	5
		30/70	0.1	0	5	10	10	16
			0.3	0	1	0	0	1
	150	50/50	0.1	0	0	0	0	0
			0.3	0	1	2	5	7
		30/70	0.1	0	0	0	0	0
			0.3	1	0	0	0	1
30%	25	50/50	0.1	0	0	2	22	23
			0.3	0	0	1	21	22
		30/70	0.1	0	1	3	10	14
			0.3	0	0	1	2	3
	150	50/50	0.1	0	0	0	0	0
			0.3	0	0	0	0	0
		30/70	0.1	0	0	0	0	0
			0.3	1	0	1	5	7

Appendix C

Sample M*plus* Code for Continuous MMIRT Model

```
TITLE: multilevel mixture IRT model -  Estimation: continuous

DATA: FILE = data.dat;

VARIABLE:
   NAMES ARE GID GLC GTHETA PID PLC PTHETA I1-I40;
   IDVARIABLE IS PID;
   USEVARIABLES = I1-I40;
   CATEGORICAL = I1-I40;

   CLASSES = c(2);
   WITHIN =I1-I40;
   CLUSTER = GID;

ANALYSIS:
   TYPE = TWOLEVEL MIXTURE;
   ALGORITHM = INTEGRATION;
   STARTS = 3 1;
   ITERATIONS = 250;
   MIXU = ITERATIONS;
   MIXC = ITERATIONS;

MODEL:
   %WITHIN%

      %OVERALL%
      f BY I1-I40* (1);
      [f@0];
      f@1;

      %c#1%
     [I1$1-I40$1];

      %c#2%
     [I1$1-I40$1];

   %BETWEEN%

      %OVERALL%
      m BY c#1;


SAVEDATA:
   RESULTS ARE C1_mfit.dat;

   FILE is C1_output.dat;
   SAVE = FSCORES;
   SAVE = CPROBABILITIES;
```

Appendix D

Sample M*plus* Code for Discrete MMIRT Model

```
TITLE: multilevel mixture IRT model -  Estimation: GLS3

DATA: FILE = data.dat;

VARIABLE:
   NAMES ARE GID GLC GTHETA PID PLC PTHETA I1-I40;
   IDVARIABLE IS PID;
   USEVARIABLES = I1-I40;
   CATEGORICAL = I1-I40;

   CLASSES = cb(3) c(2);
   BETWEEN = cb;
   WITHIN = I1-I40;
   CLUSTER = GID;

ANALYSIS:
   TYPE = TWOLEVEL MIXTURE;
   ALGORITHM = INTEGRATION;
   STARTS = 3 1;

MODEL:
   %WITHIN%
      %OVERALL%
      f BY I1-I40* (1);
      [f@0];
      f@1;

    %BETWEEN%
      %OVERALL%

      c ON cb;

MODEL c:
      %WITHIN%

      %c#1%
     [I1$1-I40$1];

      %c#2%
     [I1$1-I40$1];


SAVEDATA:
   RESULTS ARE C3_mfit.dat;

   FILE is C3_output.dat;
   SAVE = FSCORES;
   SAVE = CPROBABILITIES;
```

Bibliography

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variable regression. *Journal of Educational and Behavioral Statistics*, 22, 47–76.

Aitkin, M. (1997). The calibration of p-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood. *Statistics and Computing*, 7, 253–261.

Akaike, H. (1974). A new look at statistical model identification. IEEE *Transactions on automatic Control*, AU-19,719–722.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52,317–332.

Allua, S. (2007). *Evaluation of single- and multilevel factor mixture model estimation*. Unpublished dissertation. University of Texas at Austin.

Asparouhov, T., & Muthén, B. (2007). Multilevel mixture model. In G. R. Hancock & K. M. Samuelsen, (Eds.). *Advances in latent variable mixture models* (pp. 25–51). Greenwich, CT: Information Age Publishing, Inc.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29, 37–66.

Barron, A. R. & Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37, 1034–1054.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*. London: Arnold.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, 9, 3–29.

Bock, R. D. (1972). Estimating item parameters and latent ability when response are scored in two or more nominal categories. *Psychometrika*, 37, 29–51.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response for multiple-choice data. *Journal of Educational and Behavioral Statistics*, 26, 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, 331–348.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52,345–370.

Bozdogan, H. (1993). Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix. In: O. Opitz, B. Lausen, and R. Klar (Eds.): *Information and Classification, Concepts, Methods and Applications*. Springer, Berlin, 40–54.

Bryk, A. S., & Raudenbush, S. W. (1992). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, 101, 147–158.

Burnham, K.P., & Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. (2nd ed.). New York: Springer-Verlag.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13,195–212.

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing*, 6, 57–79.

Cheong, Y. F., & Raudenbush, S. W. (2000). Measurement and structural models for child's problem behaviors. *Psychological Methods*, 5, 477–495.

Cho, S.-J. (2007). *A Multilevel Mixture IRT Model for DIF Analysis*. Unpublished doctoral dissertation, University of Georgia.

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with applications to DIF. *Journal of Educational and Behavioral Statistics*, 35, 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. Paper presented at the International Meeting of the Psychometric Society: The 71st annual meeting of the Psychometric Society, Montreal, Canada.

Clogg, C. C. & Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of American Statistician. Association, 79,* 762–771.

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

Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research & Practice*, 20, 225–233.

Dai, Y. (2009). *Mixture Rasch Model with Covariate.* Unpublished doctoral dissertation, University of Maryland, College Park.

De Boeck, P., Wilson, M., & Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review*, 112, 129–158.

Dias, J. G. (2006). Model selection for the binary latent class model: A Monte Carlo simulation. In *Data Science and Classification*, pp 91–99. Springer Berlin Heidelberg.

Formann, A. K. (2007). (Almost) equivalence between conditional and mixture maximum likelihood estimates for some models of the Rasch type. In M. Von Davier & C. H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models* (pp. 177–190). New York: Springer.

Fox, J. -P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, 58, 145–172.

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

Fox, J. -P., & Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, 68, 169–191.

Goldstein, H. & Browne, W. J. (2002). Multilevel factor analysis modeling using Markov Chain Monte Carlo (MCMC) estimation. In Marcoulides and Moustaki (Eds.), *Latent Variable and Latent Structure Models, (*pp 225–243). New Jersey: Lawrence Erlbaum.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215–231.

Grilli, L., & Rampichini, C. (2007). Multilevel factor models for ordinal variables, *Structural Equation Modeling*, 14, 1–25.

Haertel, E. H. (1990). Continuous and discrete latent structure models for item response data. *Psychometrika*, 55, 477–494.

Hamaker, E. L., Van Hattum, P., Kuiper, R. M. & Hoijtink, H. (2011). Model selection based on information criteria in multilevel modeling. In K. Roberts & J. Hox (Eds). *Handbook of Advanced Multilevel Analysis*, pp. 231-255. Taylor and Francis.

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

Harris, D. & Sass, T. (2007). *Teacher Training, Teacher Quality, and Student Achievement.* National Center for the Analysis of Longitudinal Data in Education Research (CALDER). Working Paper #3. Washington, DC: Urban Institute.

Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. Annals of Statistics, 16, 342–355.

Hedeker, D. (1999). MIXNO: a computer program for mixed-effects nominal logistic regression. *Journal of Statistical Software*, 4, 1–92.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine*, 22, 1433–1446.

Hedeker, D. (2008). Multilevel models for ordinal and nominal variables. In J. de Leeuw & E. Meijer (Eds.), *Handbook of Multilevel Analysis*. New York: Springer.

Heinen, T. (1996). *Latent classes and discrete latent trait models*. Thousand Oaks, CA: Sage Publications.

Henry, K. L., & Muthén, B. (2010). Multilevel latent class analysis: an application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling*, 17, 193–215.

Henson, J. M., Reise, S. P., & Kim, K. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model-fit statistics. *Structural Equation Modeling*, 14, 202–226.

Hurvich, C. M. & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297–307.

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

Kane, T., Rockoff, J., & Staiger, D. (2006).*What does certification tell us about teacher effectiveness?* Evidence from New York City. Unpublished manuscript.

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, 307–327.

Landis, J.R.; & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Leroux, B. (1992), Consistent estimation of mixing distributions, *Annals of Statistics*, 20, 1350–1360.

Lissitz, R. W. (Ed.). (2005).*Value added models in education: Theory and applications*. Maple Grove, MN: JAM Press.

Lissitz, R. W. (2012) The evaluation of teacher and school effectiveness using growth models and value added models: Hope versus reality. Vancouver, BC: AERA, Division H invited address.

Lubke, G. & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: resolution by maximum likelihood. Multivariate Behavioral Research, 41,499–532.

Lukociene, O., & Vermunt, J. K (2010). Determining the number of components in mixture models for hierarchical data. In: Fink, A., Lausen, B., Seidel, W. and Ultsch, A. (eds.), *Advances in data analysis, data handling and business intelligence*, (pp 241–249). Springer: Berlin-Heidelberg.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, 26, 307–330.

Maier, K. S. (2002). Modeling incomplete scaled questionnaire data with a partial credit hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, 27, 271–289.

Maraun, M. D., Slaney, K., & Goddyn, L. (2003). An analysis of Meehl's MAXCOV-HITMAX procedure for the case of dichotomous indicators. *Multivariate Behavioral Research*, 38, 81–112

Markon, K. E. & Krueger, R. F. (2006). Information-theoretic latent distribution modeling: Distinguishing discrete and continuous latent variable models. P*sychological Methods*, 11, 228–243.

McCaffrey, D., Lockwood, J., Koretz, D., & Hamilton, L. (2003).Evaluating value-added models for teacher accountability (MG-158-EDU). Santa Monica, CA: RAND.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Miller, M. B. (1996). Limitations of Meehl's MAXCOV-HITMAX procedure. *American Psychologist*, 51,554–556.

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

Molenaar, P. C. M., & von Eye, A. (1994). On the arbitrary nature of latent variables. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 226–242). Thousand Oaks, CA: Sage.

Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345–368). Newbury Park, CA: Sage Publications.

Muthén, B. O. & Asparouhov, T. (2009). Multilevel regression mixture analysis. *Journal of the Royal Statistical Society*, Series A, 172, 639–657.

Muthén, L. K. & Muthén, B. O. (2006). M*plus* [Computer program]. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55,463–469.

Palardy, G., & Rumberger, R. W. (2008). Teacher Effectiveness in First Grade: The Importance of Background Qualifications, Attitudes, and Instructional Practices for Student Learning. *Educational Evaluation and Policy Analysis*, 30, 111–140.

Palardy, G., & Vermunt, J.K. (2010). Multilevel growth mixture models for classifying groups. *Journal of Educational and Behavioral Statistics,* 35, 532–565.

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.

Raudenbush, S. W., & Bryk, A. G. (2002). H*ierarchical Linear Models: Applications and data analysis methods* (2nd eds.), Thousand Oaks, CA: Sage.

Reise, S. P., & Gomel, J. N. (1995). Modeling qualitative variation within latent trait dimensions: Application of mixed-measurement to personality assessment. *Multivariate Behavioral Research*, 30,341–358.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica,* 14, 465-471.

Rodr ́guez, G., & Elo, I. (2003). Intra-class correlation in random-effects models for binary data. *The Stata Journal,* 3, 32–46.

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

Rost, J. (1997). Logistic mixture models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449–463). New York: Springer.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record*, 104, 1525–1567.

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective*. Unpublished doctoral dissertation, College Park: University of Maryland.

Sanders, W., Saxton, A., & Horn, B. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluational measure?* (pp. 137–162). Thousand Oaks, CA: Corwin.

Schacter, J., & Thum, Y. M. (2004). Paying for high and low-quality teaching. Economics of Education Review, 23, 411–430.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464

Sclove, S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333–343.

Skrondal, A., & Rabe-Hesketh, S. (2004).*Generalized latent variable modeling*. Boca Raton, FL: Chapman & Hall.

Smit, A., Kelderman, H., & Flier, H. (1999). Collateral information and mixture Rasch models. *Methods of Psychological Research Online*, 4.

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

Van Horn, M. L., Fagan, A. A., Jaki, T., Brown, E. C., Hawkins, J. D., Arthur, M. W., et al. (2008). Using multilevel mixtures to evaluate intervention effects in group randomized trials. *Multivariate Behavioral Research*, 43, 289–326.

Varriale, R., & Vermunt, J. K. (2012). Multilevel mixture factor models. *Multivariate Behavioral Research*, 47, 247–275.

Vermunt, J. K. (2001) The use restricted latent class models for defining and testing nonparametric and parametric IRT models. *Applied Psychological Measurement*, 25, 283–294.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, 33, 213–239.

Vermunt, J. K. (2007). A hierarchical mixture model for clustering three-way data sets. *Computational Statistics and Data Analysis*, 51, 5368–5376.

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

Vermunt, J. K. (2008b). Multilevel latent variable modeling: An application in education testing. *Austrian Journal of Statistics*, 37, 285–299.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18, 450–469.

Vermunt, J. K., & Magidson, J. (2005). Technical guide for Latent GOLD 4.0: Basic and advanced. Belmont MA: Statistical Innovations Inc.

Vermunt, J. K. & Magidson, J. (2008). LG-Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module, Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K. & Van Dijk. L. (2001). A nonparametric random-coefficients approach: the latent class regression model. *Multilevel Modeling Newsletter*, 13, 6–13.

Weaklim, D. L. (2004). Introduction to the special issue on model selection. *Sociological Methods and Review*, 11, 192-196.

Wong, G. Y. & W. M. Mason (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80, 513–524.

Wu, M. L., Adams, R. L., & Wilson, M. R. (1997).Conquest: Generalized item response modeling [Computer software]. Victoria: Australian Council for Educational Research.

Yang, C. (2006). Evaluating latent class analyses in qualitative phenotype identification. *Computational Statistics Data Analysis*, 50, 1090–1104.