A Multilevel Mixture IRT Model for DIF Analysis

by

Sun-Joo Cho

(Under the direction of Allan S. Cohen)

Abstract

The usual methodology for detection of differential item functioning (DIF) is to examine differences among manifest groups formed by such characteristics as gender, ethnicity, age, etc. Unfortunately, membership in a manifest group is often only modestly related to the actual cause(s) of DIF. Mixture item response theory (IRT) models have been suggested as an alternative methodology to identifying groups formed along the nuisance dimension(s) assumed to be the actual cause(s) of DIF. A multilevel mixture IRT model (MMixIRTM) is described that enables simultaneous detection of DIF at both examinee- and school-levels. The MMixIRTM can be viewed as a combination of an IRT model, an unrestricted latent class model, and a multilevel model. Three perspectives on this model were presented: First, the MMixIRTM can be formed by incorporating mixtures into a multilevel IRT model; second, the MMixIRTM can be formed by incorporating a multilevel structure into a mixture IRT model; and third, the model can be formed by including an IRT model in a multilevel unrestricted latent class model. A fully Bayesian estimation of the MMixIRTM was described including analysis of label switching, use of priors, and model selection strategies along with a discussion of scale linkage. A simulation study and a real data example were presented.

Index words: Bayesian Estimation, Differential Item Functioning, Finite Mixture Modeling, Item Response Theory, and Multilevel Modeling

A Multilevel Mixture IRT Model for DIF Analysis

by

Sun-Joo Cho

B.A., Yonsei University, 1999

B.A., Yonsei University, 2001

M.S., Yonsei University, 2003

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2007

A Multilevel Mixture IRT Model for DIF Analysis

by

Sun-Joo Cho

Approved:

Major Professor: Allan S. Cohen

Committee: Daniel B. Hall
Deborah L. Bandalos
Seock-Ho Kim

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2007

Dedication

To mom and dad

my teachers, Professors Allan Cohen, Sang-Jin Kang, and Seock-Ho Kim

Acknowledgments

I would like to sincerely thank my advisor, Professor Allan Cohen, for his guidance, understanding, patience, and supports for my doctoral study. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own research interests, and at the same time all kinds of supports. And my deepest gratitude is to Professor Seock-Ho Kim. He has been always there to listen and give me advice. I am deeply grateful to him for the long discussions that helped me sort out technical details of my works. I also gratefully acknowledge Professor Deborah Bandalos who has given her time to read my manuscript and also offered valuable advice during my graduate program at the University of Georgia. I would also like to thank Professor Sang-Jin Kang at Yonsei University for his encouragement and advice.

My thanks go to the College Board, for the grant they awarded me for this study. I also would like to thank Dr. Shan-ho Tsai at the Research Computing Center, University of Georgia, for her help to use the Linux cluster even during a winter break.

Finally, this dissertation would not have been possible without love and supports of my mom and dad. I would also like to thank my friends, especially Ji and Young-Suk, for their love and encouragement. They were always with me at the best and worst moments of my dissertation journey. And I owe a special note of gratitude to somebody who makes me feel happy with his mails.

CUPS, Athens, GA Sun-Joo Cho

September, 2007

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Statement of Problem

The Preliminary Scholastic Aptitude Test (PSAT)/National Merit Scholarship Qualifying Test (NMSQT) is a standardized testing program co-sponsored by the College Board and the National Merit Scholarship Corporation. The program provides firsthand practice for the Scholastic Assessment Test (SAT) Reasoning Test to high school students considering taking the SAT. A major concern of the College Board is providing feedback to schools of students who have participated in this program.

The Summary of Answers and Skills (SOAS) is a report provided by the College Board to each school that has 25 or more students participating in the PSAT/NMSQT program. The SOAS gives a snapshot of a given school's performance on the PSAT/NMSQT items administered in a given year. The report allows schools to compare their students' performance with those of students in what is termed a comparable group of schools as well as with state and national groups. One aspect of the report compares the performance on an item by item basis of students in the given school with those in the comparision group. This part of the SOAS report is based on a differential item functioning (DIF) analysis in which a standardized p-difference (STD P-DIF) analysis is used to compare the item performance in the given school to that in the comparision group after matching on ability as measured by the PSAT/NMSQT. The SOAS report provides extensive information about item performance of the given school relative to a comparision group defined based on schools with a similar performance profile.

A comparision group is a statistically created group of schools with this same performance profile and used for illustrating what the expected performance on each item would have been of schools like the given school. Feedback from schools indicates that the concept of how the comparable group is defined is difficult to explain. Further, schools have indicated they would like greater flexibly in comparing the performance of their students with those from other schools that have similar demographic characteristics in order to better understand why their students may or may not have performed as well as students in the comparision group.

1.2 The Purpose of the Study

In this study, we examine an alternative approach to defining comparison groups and provide a means of detecting items which function differentially. DIF is said to arise, when different groups of examinees, who are of the same ability, have different probabilities of getting a question correct (Pine, 1977). More generally, DIF arises when the propensity for a particular response differs among groups of examinees conditioned on the ability being measured by the test. The presence of DIF is a serious problem in educational testing as it indicates a threat to the validity of the test (Thissen, Steinberg, & Wainer, 1988, 1993). The purpose of this study is to develop a general methodology that identifies and describes characteristics of comparison groups in the context of a DIF detection analysis. In so doing, the method provides a descriptive profile of schools that are like the given school based on easily observed and understood characteristics. This study develops a model that integrates an IRT model, a finite mixture model (i.e., a latent class model), and a multilevel model. In addition, this study presents a method of estimating model parameters with a Bayesian solution.

1.3 Significance of the Study

DIF arises when groups differ in item level performance on one or more nuisance dimensions after conditioning on some matching variable (Ackerman, 1992; Roussos & Stout, 1996). Use of manifest groups to explain that performance differential, however, does not accurately

reflect the actual causes of the DIF (DeAyala et al., 2002; Cohen & Bolt, 2005; Cohen, Gregg, & Deng, 2005; Samuelsen, 2005). The causes of DIF may be more accurately assessed by analyzing performance along the nuisance dimension(s) rather than focusing on the characteristics associated with manifest group membership such as gender or ethnicity. Previous research suggests that differences among latent classes may be due, for example, to differences in curricular experiences (Cohen, Gregg, & Deng, 2005) or to difficulty with particular types of items rather than with membership in a particular manifest group (Kang & Cohen, in press). In the general approach described in this study, latent classes are formed on the basis of homogeneities among class members with respect to response patterns. Members of these classes differ in ways that are directly related to the response strategies used for answering questions on the test (Mislevy & Verhelst, 1990; Rost, 1990).

An additional component addressed in the model developed in this study focuses on the impact of the hierarchical structure that is typical of much educational and psychological data such as those from the PSAT/NMSQT. If this structure is not accounted for in the model, the model used to scale the item responses will be inadequate. One result of this is that detection of the latent classes will likely be obscured. The model proposed for this study will account for this structure using a multilevel model combined with a mixture IRT model (MixIRTM). This approach addresses a general problem in the use of standard DIF detection methods.

DIF detection methods that focus on manifest group membership assume a homogeneity among group members that is not always justified. Use of the proposed model in this study accounts for the natural nesting that is characteristic of much educational data, and also provides a means of explaining why the differences among latent classes may have arisen by obtaining the homogeneous group (i.e., latent class) with respect to their response pattern. Such information has greater explanatory power than the standard DIF test and may provide users such as the College Board with a tool to help schools more easily understand how they are like schools in the same latent class and how they are dissimilar from schools in other

latent classes. This explanation is expressed using easy to understand information about the kinds of questions that are differentially harder or easier for one latent class compared to others and the characteristics of examinees and of schools that are members of these latent classes. Providing this type of information may help in conceptualizing the reasons items are functioning differentially. It also may even be useful for revising the items, as item writers can incorporate into any revisions the fact that the item functions differently for schools of a particular latent class.

1.4 Overview of Chapters

The first section of Chapter 2 provides a description of the STD P-DIF method currently used in the PSAT/NMSQT program. Next, the general rationale for DIF analysis using the MixIRTM is described, followed by the description of the proposed model.

The multilevel mixture IRT model (MMixIRTM) can be viewed as a combination of an item response theory (IRT) model, a finite mixture model (or latent class model), and a multilevel model. We show the development of this model as a combination of three separate perspectives on modeling item response data using a multilevel mixture modeling approach. First, the MMixIRTM is considered as an extension of multilevel IRT models and is described to include individual-level and school-level mixtures. The multilevel structure of test data is described using this perspective followed by the incorporation of the multilevel structure into IRT. In the second perspective, the MMixIRTM is described as an extension of a MixIRTM in order to include the multilevel structure of test data. A finite mixture model is described in this section and then the incorporation of the mixture model into an IRT model is presented. In the third perspective, the MMixIRTM is described as an extension of a multilevel latent class model. This is described as an unrestricted latent class model and used to show this third perspective indicating how the IRT model can be included in the context of a latent class model. Finally, the MMixIRTM as a general model and its three special models are

described as the extension of the three perspectives. In addition, covariates are incorporated into the MMixIRTM and use of the model for DIF detection is described.

Chapter 3 describes some Bayesian estimation issues that are applicable for both multilevel IRT and MixIRTMs. A fully Bayesian estimation of the MMixIRTM is presented along with a simulation study designed to assess the recovery of the proposed model. Chapter 4 presents the results of the simulation study. The proposed model was also applied to the PSAT/NMSQT Mathematics test to illustrate the proposed model and then that result was compared with results from the STD P-DIF analysis. Chapter 5 summarizes the methods and results, and discusses limitations and possible future work.

Chapter 2

Theoretical Background

The presence of DIF is an indication that the items on the test do not measure the latent variable that is the intended focus of the test (Ackerman, 1992). To address this, the development and application of a MMixIRTM is described for use in detection of DIF. The multilevel portion of the MMixIRTM is used to account for the inherent nesting of students within schools that is present in much educational testing data, such as data from the PSAT/NMSQT. Ignoring this hierarchical data structure can lead to distorted estimates of model parameters, particularly random components such as ability (Adams, Wilson, & Wu, 1999). In this study, we present a DIF detection model using the MMixIRTM that accounts for this natural hierarchical structure at both a student-level and at a school-level.

The mixture portion of the MMixIRTM is used to identify latent groups (or classes) at two levels, the student-level and the school-level. At the student-level, the latent classes are composed of persons who are homogeneous in their use of particular response strategies (Mislevy & Verhelst, 1990; Rost, 1990). At the school-level, the latent classes are composed of schools which share similar characteristics (Vermunt & Magidson, 2005). The MMixIRTM is extended at the person-level to incorporate person-level covariates such as gender, ethnicity, socio-economic status. At the school-level, the model is extended to include school-level covariates, such as socio-economic status and urban/suburban/rural location, to describe the composition of the latent groups at each level. This model incorporates an approach to DIF analysis based on differences in examinee response patterns described by Cohen and Bolt (2005).

In addition, the IRT model portion of the MMixIRTM provides item-level information that can be used to identify questions which are differentially harder or easier for students and for schools in each latent group. The complete model is described below, preceded by a description of the STD P-DIF method.

The STD P-DIF method currently used in the PSAT/NMSQT program is described, followed by the description of the proposed model in this chapter. Three perspectives on this model are presented below: First, the MMixIRTM can be formed by incorporating mixtures into a multilevel IRT model; second, the MMixIRTM can be formed by incorporating a multilevel structure into a MixIRTM; and third, the model can be formed by including an IRT model in a multilevel unrestricted latent class model.

2.1 Standardized P-DIF

The STD P-DIF statistic (Dorans & Kulick, 1986) is currently used by the PSAT/NMSQT program to compare the item performance in a particular school (the focal school) to that in a reference group population of schools after matching on ability (i.e., the total score). Dorans and Kulick (1986) describe the STD P-DIF as

$$STDP-DIF = \sum_{s=1}^{S}[\frac{K_s}{\sum_{s=1}^{S} K_s} \cdot (P_{fs} - P_{rs})], \tag{2.1}$$

where s refers to the score interval of the matching variable, $\frac{K_s}{\sum_{s=1}^{S} K_s}$ is the weighting factor at score level s supplied by the standardized group and used to weight differences in performance between the focal group P_{fs} and the reference group P_{rs}. STD P-DIF weights the difference in terms of a numerical index indicating the discrepancy between the reference and focal groups. This approach emphasizes the importance of comparing groups based on a matching variable. In practice, the matching variable is the total test score.

As it is implemented, the STD P-DIF is defined as the difference in item performance between the focal group and reference group members matched on the appropriate PSAT/NMSQT score. The calculation of STD P-DIF for PSAT/NMSQT is same as that

used by Dorans and Kulick (1986). The proportion of students in each score level supplied by the focal group is given as $\frac{K_s}{\sum_{s=1}^{S} K_s}$. As indicated above, in the context of the PSAT/NMSQT analysis provided in the SOAS, the comparision group is defined as the reference group and the local group is defined as the focal group.

The STD P-DIF can take values from -100% to 100%. Positive values indicate the studied item favors the focal group; negative values indicate the studied item favors the reference group. Currently, the College Board views STD P-DIF values between -5% and 5% as containing negligible DIF, although STD P-DIF values between -10% and -5% and between 5% and 10% are inspected to ensure that no possible effect is overlooked. Items with STD P-DIF values outside the $-10\%, 10\%$ range are more unusual and are examined very carefully.

2.2 Differential Item Functioning Analysis with Mixture IRT Models

Mixture Models for DIF Analysis. DIF analysis is typically based on manifest grouping variables such as gender or ethnicity. This approach to DIF analysis makes an implicit assumption that the characteristic that is the cause of the DIF is homogeneous within a given manifest group. Unfortunately, this is not usually the case. This approach differs, however, from a more recently developed one in which DIF is modeled as a function of nuisance dimensionality not accounted for directly by the model. The nuisance dimensionality is assumed to be the actual cause(s) of the DIF (DeAyala et al., 2002; Cohen & Bolt, 2005; Cohen, Gregg, & Deng, 2005). As Samuelsen (2005) has shown, this dimensionality is only moderately related at best to the manifest variables (e.g., gender or ethnicity) commonly assumed to be the cause of DIF. Once a DIF item has been identified using the standard approach for DIF detection, little is known about the examinees for whom the item functions differentially. This is because DIF detection methods focus on manifest group characteristics that are associated with, but do not explain why examinees respond differentially to items.

Samuelsen (2005) notes three problems with respect to the detection of DIF using simply a manifest grouping variable. First, the manifest grouping variables often do not represent homogeneous populations. The Hispanic population in the United States, for example, is diverse in origin and ethnicity, such that classification as Hispanic does not yield a group that is homogeneous on a dimension related to DIF. With respect to DIF at a school-level, von Davier and Yamamoto (2004) illustrated this by suggesting that the data can be viewed as partially missing, because traditional school types are a track system with students ranked in proficiency, whereas the mixed-school types generally accept a broader range of students. Concrete examples include types of schools in which some set the standard of comparisons, perhaps for historic reasons, such as the well-known public and private. Other schools (e.g., schools based on some faith or religion or on a special educational theory) may have innovative concepts that make it impossible to classify them as one of the traditional types. This setup is often due to ambiguous responses that do not allow grouping all schools into a fixed number of known categories. In addition, not all schools offer the same opportunities to their students; there are differences in staff, equipment, and course offerings (O'neill & Mcpeek, 1993). Another example is the recent reform in marking ethnicity in large-scale assessments; this variable is now allowed to have multiple responses, so students can indicate more than one ethic group to which they belong. Reference and focal groups based on ethnicity and gender are quite heterogeneous and their difference are not easy to describe (Schmitt, Holland, & Dorans, 1993).

The second problem described by Samuelsen (2005) is that the groups being affected by DIF are not well-defined by the manifest groups being studied. Hu and Dorans (1989) note that the removal of an item favoring females resulted in slightly lower scores for females and slightly higher scores for males (Samuelsen, 2005). Scores of both Hispanics and Asian-Americans, however, were subsequently raised more than scores of males, meaning that females in those groups actually received an advantage from the removal of the item favoring females.

The third problem resulting from use of a manifest grouping variable is that the manifest groups used for DIF comparisons are not directly related to the issues of learning about which educators care. Manifest groups defined by characteristics such as gender and ethnicity are really proxies for something else. As a result, using manifest groups for DIF comparisons instead of the real sources of the nuisance dimensionality leaves open the clear possibility that we may actually miss items that are functioning differentially based on the latent attribute(s). In addition, it is incorrect to assume an item or set of items exhibiting DIF disadvantages all members of a manifest group. Finally, for items that exhibit DIF, the true magnitude of the DIF may be obscured due to the lack of overlap between the manifest groups and the latent classes.

Samuelsen (2005) notes that the use of manifest groups will result in inflated Type I error rates in detection of DIF, loss of power, and underestimation of the magnitude of the DIF, when manifest groups and latent classes do not completely overlap. As the amount of overlap between manifest groups and latent classes decreases, however, so will the power to correctly identify items with DIF. Further, as overlap decreases, the true magnitude of the DIF is increasingly obscured, making it more difficult to successfully detect items.

Perspectives of DIF Analysis with Mixture IRT model. The presence of different response patterns on a test may be an indication of the use of different response strategies. Depending on the purpose of the test, this may or may not be a problem. When decisions that rely on uses of different strategies are of interest, however, useful information is less likely to be obtained, when these different response patterns are ignored. This is the usual case with existing conventional tests, which are currently often constructed so that differential patterns of responding are minimized or at least ignored in the analysis. In fact, most current academic achievement tests are constructed to minimize differentiation among strategies, since such differences act to lower the reliability of overall scores (Mislevy & Verhelst, 1990). When this is the focus of test construction, different response strategies may be construed as evidence of lack of construct validity. The impetus for using MixIRTM is that we assume the population

of examinees consists of latent classes defined by important differences in use of response strategies for answering items on a test.

MixIRTMs have been used to detect latent subpopulations that differ systematically in their responses on educational tests. Rost (1990, 1997) used a mixture Rasch model (MRM) to detect qualitative differences among examinees. A mixture linear logistic test model was used to detect random guessing behavior on a multiple-choice test (Mislevy & Verhelst, 1990). Bolt, Cohen, and Wollack (2001) similarly used a mixture nominal model to investigate individual differences in the selection of response categories in multiple-choice items. A MRM with ordinal constraints was used to model test speededness (Bolt, Cohen, & Wollack, 2002) and to help maintain scale stability in the presence of test speededness (Wollack, Cohen, & Wells, 2003). MixIRTMs have also been used for detecting differential functioning on items (Cohen & Bolt, 2005; Cohen, Gregg, & Deng, 2005; Samuelsen, 2005), a testlet (Cohen, Cho, & Kim, 2005), and suggested an item bundle or a scale (von Davier & Yamamoto, 2004). This methodology enables the identification of subgroups for which an item or a group of items function differently among groups without the need of specifying these groups priori. As Mislevy and Verhelst (1990) and Rost (1990) have noted, this is done by fitting an IRT model to the different latent classes, each of which differs in the propensity to answer the items on the test. This is a key element in the rationale for the development of MMixIRTM in the present study.

Dimensionality is not a function of the test by itself, but a function of the test in the context of a particular group of examinees (Angoff, 1993). Although the MixIRTM technically contains only one ability level for each person, it can be viewed as multidimensional because (1) the probability of a particular response is predicted with class membership, and (2) the order of item difficulty varies across persons, as in multidimensional models. The MixIRTM approach is consistent with Ackerman (1992) and with the multidimensional DIF analysis described by Roussos and Stout (1996), in that both deal with residual variation not accounted for by unidimensional IRT models. The two approaches are different, however, in

the following ways. First, Ackerman (1992) and Roussos and Stout (1996) identify construct-related dimensions that elicit group differences with the multidimensional IRT model. The MixIRTM approach, on the other hand, follows the method of Cohen and Bolt (2005) that identifies construct-related categories with the mixture portion of the MixIRTM.

Being dimension-like requires within category homogeneity and between-category quantitative differences and being category-like requires within-category homogeneity and between-category qualitative differences (De Boeck et al., 2005). As Ackerman (1992) illustrated, nuisance abilities can be thought of as skills which the examinee uses to solve particular items but which were not intended to be assessed. In his example, reading ability may be considered to be a nuisance skill in a test designed to measure the pure ability of algebraic symbol manipulation. Rost (1990) identifies two latent classes (i.e., categories) that reflected knowledge states on a physics achievement test. For one class, items that concerned textbook knowledge were relatively more difficult than items that reflected practical experience. For the other class, the opposite pattern was observed.

Ackerman (1992) and Roussos and Stout (1996) suggest that the nuisance dimensionality is not accounted for by a unidimensional IRT model and, as a result, is the cause of DIF. DIF, from this perspective, is related to the presence of one or more secondary dimensions not accounted for by the primary construct of interest. The Cohen and Bolt (2005) approach addresses the issue of residual variation by seeking to isolate unknown or latent groups of persons. This can be done in an exploratory or a confirmatory fashion. The MixIRTM is used to identify latent classes of examinees which are homogeneous with respect to item response patterns. The members of each latent class may vary in ability, but the response strategies differ among classes.

Embretson and Reise (2000) suggest two ways in which identifying latent classes is important for understanding validity. First, if there is only one class, then it is correct to assume that all examinees use the same strategies. If two or more latent classes are present, however, the nature of the construct depends on the characteristics of the examinees in each latent

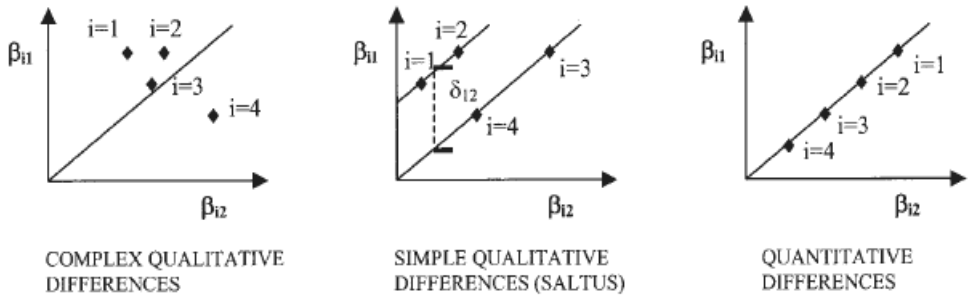
class. Class membership actually may be a moderator variable, in fact, in the sense that it may be more important for criterion-related validity for one latent class than another. The second way in which latent classes are important is in the way in which comparisons are made among groups. The typical DIF analysis, including that for Ackerman (1992) and Roussos and Stout (1996) is based on manifest groups. Items that measure the secondary dimension and produce DIF demonstrate a quantitative difference between the manifest focal and reference group(s) due to the nuisance dimension(s). However, the DIF analysis described by Cohen and Bolt (2005) is based on the comparisons among two or more latent classes.

It is often the case in performing a DIF analysis that comparisons are made between a single reference group and multiple focal groups (Penfield, 2001). It is also possible that multiple reference groups are possible. It is this latter approach, that is, the possibility of multiple reference groups, that is most useful for the College Board PSAT/NMSQT program as that program seeks to provide individual schools with information about the performance of their particular school relative to other groups of schools.

Patterns of Mixtures. Rost (1990) suggested the primary diagnostic potential of the MRM (i.e., without school-level mixtures) is in its use for accounting for qualitative differences among examinees. The presence of nonintersecting item parameter profiles justifies the interpretation that there are quantitative differences between abilities among the different latent classes. Similarly, intersecting item parameter profiles indicate the presence of qualitative differences. De Boeck et al. (2005) describe two types of qualitative differences between manifest categories. These descriptions also can be applied to differences among latent categories. The qualitative differences are simple when item difficulties in one latent class have shifted relative to the item difficulties of the other latent class(-es). The qualitative differences are complex when there does not appear to be a discernible pattern to the item difficulties among the latent classes. When qualitative differences are not present between two categories, then location (and possibly discrimination) are the same in the two categories. This is shown in the right panel in Figure 2.1 (reused from De Boeck et al, 2005).

If item difficulties are same in the two categories, then quantitative differences are present. Quantitative and qualitative differences are shown in Figure 2.1.

Figure 2.1: Different Kinds of Differences, reused from De Boeck et al. (2005)



Description of Latent Classes. DIF analysis with MixIRTMs has been shown to be useful for helping to understand the causes of DIF, although with respect to latent populations of examinees rather than with manifest ones. By classifying individuals into latent classes, we then focus efforts on describing (and measuring) the members of each class with respect to specific attributes. Typically, in an educational testing context, we do this by examining possible attributes of educational advantage or disadvantage. In this study, two procedures were used to describe the latent classes. First, it is possible to compare the membership of examinees classified into different latent classes (i.e., the estimated group membership) with their manifest group memberships and then do some type of association analysis. A second method is to model mixtures with covariates that help to describe members of each latent class.

The use of covariates also has been shown to improve detection of the latent classes (Smit, Kelderman, & van der Flier, 1999; Cho, Cohen, & Kim, 2006). The covariates that are selected are those that are of potential help in describing the characteristics of mixtures.

In the context of this study, this is going to be important at both student- and school-levels. In other words, the description of mixtures with covariates helps to describe which students are clustered with respect to both student- and school-level mixtures. Student- and school-level covariates will be selected from variables in the PSAT/NMSQT dataset. In addition, the characteristics of mixtures can be described with respect to the content and cognitive skills required for each item by examining the class-specific item difficulties.

A Strategy of Use of Mixture IRT DIF Procedure. Samuelsen (2005) suggested a four-step approach for examining DIF using a latent class perspective. The first step is to assess how many mixtures fit the data based on both statistical model selection criteria and on substantive rationale. The second step is to determine the appropriateness of the use of the manifest group as a proxy for the latent group. That is, it is necessary to determine to what extent the manifest group overlaps with the latent group. Samuelsen (2005) suggested it would be appropriate to indicate that an item functioned differentially against one manifest group if the overlap with the latent group was 99 percent. It would be inappropriate, however, to use the manifest groups for DIF detection if the overlap was only 60 percent. In a two-group model, this would mean 40 percent of the people have response strategies like those in the other manifest group. The third step is to examinee the data from the latent class analysis for clues as to why the items may have functioned differentially. Samuelsen suggests starting by examining mean abilities within the latent classes, differences in item difficulties between the latent classes, and patterns of item difficulties within classes. The fourth step is to use covariates to predict latent class membership.

Comparison of STD P-DIF and MMixIRTM for DIF Detection. STD P-DIF is based on the observed item scores uses a manifest grouping variable to define the focal and reference groups. The proposed model for DIF analysis in this study, however, is based on using a latent variable, as measured by an IRT model, and a latent grouping variable, as indicated by latent class membership.

The comparision group for STD P-DIF analysis is composed of students who have critical reading, mathematics, and writing scores between 20 and 80 inclusive in Grades 10 and 11 for the current school year. Non-standard students are excluded. Formation of comparision groups with the MMixIRTM approach, however, is done based on students' item response patterns.

2.3 Latent Variable Modeling Approaches

In this section, different modeling approaches are described for representing the latent variable (i.e., the latent ability) in the MMixIRTM. Each of the modeling approaches is presented in order to help distinguish among some of the possible approaches that can be used to model the latent variable. The intention is to provide a context for the approach taken in this study in developing the MMixIRTM.

IRT, latent class, and multilevel models each have their own features as shown in Table 2.1. IRT models have an advantage in that they include the invariance property and explicit functional relationships between response probabilities and the latent variable(s). The unrestricted latent class model (LCM), likewise, is useful in that it is often used for the clustering, that is, for finding homogeneous groups among the data. Multilevel models also can be developed to account for the multilevel structure that is present in so much educational testing data.

The three models are combined in Table 2.1 to illustrate how features of each of the models fit together. The MixIRTM itself is a combination of an IRT model and a LCM. Multilevel IRT is a combination of an IRT model and multilevel models. Multilevel latent class model (MLCM) is a combination of multilevel models and LCM. The combined models do have some flaws as noted in Table 2.1.

The proposed MMixIRTM is described in the sequel as a combination of three different models: a MixIRTM, a multilevel IRT model, and a multilevel LCM. The three different combinations lead to three different perspectives for the MMixIRTM as shown in the Venn

Table 2.1: Features of Models Combined to Form the Multilevel Mixture IRT Model

Model	Unique Features	Aspects of Flaws
1. IRT	Invariance property Explicit functional relationship between response probability and latent variable	Mixture, Multilevel
2. (Unrestricted) LCM	Data reduction/Clustering	IRT, Multilevel
3. Multilevel	Incoporation of multilevel structure	IRT, LC
4. MixIRTM		Multilevel
5. Multilevel IRT		Mixture
6. MLCM		IRT
7. MMixIRTM		None

diagrams in Figure 2.2: First, a MMixIRTM can be formed by the incorporation of mixtures into a multilevel IRT model; second, a MMixIRTM can be formed by incorporating a multilevel structure into a mixture IRT model; and third, a MMixIRTM can be composed of an IRT model included in a multilevel unrestricted latent class model.

IRT Model. The latent variable, ability (θ_j), is typically modeled in IRT as follows:

$$\theta_j \sim N(0, 1), \tag{2.2}$$

where $j = 1, \ldots, J$ students.

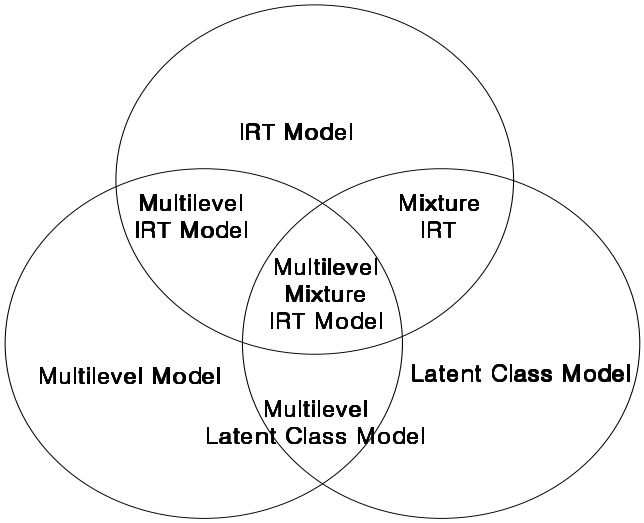
Latent Class Model. The latent variable in a LCM is as follows:

$$\theta_j = \sum_{g=1}^{G} \pi_g \cdot \theta_g, \tag{2.3}$$

where θ_g ability is discrete in a latent class, and π_g is the proportion of students in class g. In this study, a restricted form of the LCM is of interest.

Multilevel Model. There is no latent variable in the multilevel model for incorporating measurement error. However, the random effects in the multilevel model are arbitrarily assumed to be normally distributed in the same way as ability is assumed to be distributed

Figure 2.2: Latent Variable Modeling



in IRT model. Combining the two models yields a multilevel IRT model. When the latent variable is assumed to be discrete, as is the case for the LCM, the result is a multilevel LCM. Thus, there is some similarity between multilevel modeling and latent variable modeling. This similarity can lead to combined models like a multilevel IRT model, which is a combination of a multilevel nonlinear model and an IRT model, and multilevel structural equation modeling (SEM), which is a combination of a multilevel linear model and a SEM. The multilevel nonlinear model is of interest in this study.

Mixture IRT Model. The latent variable modeling in the MixIRTM is as follows:

$$\theta_{jg} \sim N(\mu_g, \sigma_g^2), \tag{2.4}$$

where $j = 1, \ldots, J$ students, $g = 1, \ldots, G$ student-level latent class, μ_g is the mean of ability for a class g, and σ_g^2 is the variance of ability for a class g.

Multilevel IRT Model. Modeling of the latent variable in the multilevel IRT model is as follows:

$$\theta_{jt} = \gamma_{00} + u_{jt} + \nu_t, \tag{2.5}$$

where $j = 1, \ldots, J$ students, $t = 1, \ldots, T$ schools, γ_{00} is the average ability for school t, u_{jt} is the ability across students, u_{jt} follows $N(0, \tau)$, which τ is the variance of ability at the student level, and ν_t follows $N(0, \zeta)$, which ζ is the variance of ability at the school level. The total variance of ability in the multilevel IRT model is decomposed into the student-level ability variance (i.e., τ) and the school-level ability variance (i.e., ζ).

Multilevel Latent Class Model. The latent variable modeling in a MLCM is as follows:

$$\theta_{jt} = \sum_{k=1}^{K} \sum_{g=1}^{G} \pi_k \cdot \pi_{g|k} \cdot \theta_{gk}, \tag{2.6}$$

where $j = 1, \ldots, J$ students, $t = 1, \ldots, T$ schools, $g = 1, \ldots, G$ student-level latent classes, and $k = 1, \ldots, K$ school-level latent classes. If the restricted LCM is used, this is a non-parametric multilevel IRT model.

Multilevel Mixture IRT Model. Modeling of the latent variable in the MMixIRTM developed in this study is as follows:

$$\theta_{jtgk} \sim N(\mu_{gk}, \sigma_{gk}^2), \tag{2.7}$$

where $j = 1, \ldots, J$ students, $t = 1, \ldots, T$ schools, $g = 1, \ldots, G$ student-level latent classes, $k = 1, \ldots, K$ school-level latent classes, μ_{gk} is the mean of ability for classes g and k, and σ_{gk}^2 is the variance of ability for classes g and k.

2.4 The Multilevel Structure of Test Data and a Multilevel IRT model

2.4.1 The Multilevel Structure of Test Data

The basic structure of PSAT/NMSQT data is hierarchical in that item responses are nested within students and students are nested within schools. This is typical of much educational

data, and creates a data structure which presents problems for estimation of some item-level, person-level, and school-level parameters. Students within a particular school, for example, share the same physical environment, and have a set of similar experiences, including similar instructional sequencing, similar teachers, and similar peers. The assumption that observations are independent of school effects, therefore, is potentially violated. The result of this violation is that the standard errors of parameters realized in a traditional regression analysis will be often too small. This fact makes the traditional regression analyses (i.e., aggregated regression and disaggregated regression for the multilevel structure) less conservative.

Several studies have demonstrated improvements in parameter estimation for hierarchically structured data with use of multilevel analysis. With respect to estimates of fixed parameters, both traditional regression analysis and multilevel analysis have been found to produce unbiased estimates (Kim, 1990; Bassiri, 1988). The improvement in estimates of parameters using multilevel analysis arises for random components of the model. This is because the standard errors of parameters in multilevel modeling are not underestimated.

2.4.2 Multilevel Item Response Theory Model

Multilevel models, also known as hierarchical linear models (HLM), allow the natural multilevel structure of educational and psychological data to be represented formally in the analysis of the data (Bryk & Raudenbush, 1992; Goldstein, 1987; Longford, 1993). The combination of HLM with IRT provides more accurate estimation of the standard errors of the parameters (Adams et al., 1997; Maier, 2001, 2002; Fox, 2005). This combination also has led to the development of psychometric models for item response data that contain hierarchical structure, thus enabling a researcher to study the impact of different predictors such as schools and curriculum on the lower level units (e.g., students) (e.g., Adams et al., 1997; Kamata, 2001; Maier, 2001, 2002; Fox & Glas, 2001).

Kamata's Three-Level IRT Model

A useful model to begin with is the three-level IRT model developed by Kamata (2001). This model is explained below and then extensions are made to develop the MMixIRTM. The IRT portion of Kamata's model is a Rasch model. The multilevel IRT model formulation for binary items in that model is presented below.

First-Level Model. The first level of this model identifies the measurement model.

$$\log\left[\frac{P_i(\theta_{jt})}{1-P_i(\theta_{jt})}\right] = \theta_{jt} - \beta_{ijt} \tag{2.8}$$

with $j = 1, \ldots, J$ examinees, $t = 1, \ldots, T$ schools, for a test of $i = 1, \ldots, I$ items, θ_{jt} is the ability of examinee j in school t, and $P_i(\theta_{jt})$ is the probability of person j in a school t answering the item i correctly, and β_{ijt} is the difficulty of item i. β_{ijt} has two subscripts, j and t, even though item difficulty is actually a characteristic of the item. The reason is that at the first level, the regression coefficients in the hierarchical linear model are expressed for each item, but these are not treated as variables in the regression analysis. The j and t subscripts for items are removed at the second and third levels of the model.

Second-Level Model. The second level of the model is the student- or person-level of the model. At this level, the model provides student ability estimates as follows:

$$\theta_{jt} = \gamma_{0t} + u_{jt} \ , \tag{2.9}$$

where γ_{0t} is the average ability of individuals js in a school t, and u_{jt} follows $N(0, \tau)$, in which τ is the variance of ability at the student level. Item difficulty estimates are given as

$$\beta_{ijt} = \beta_{it} \tag{2.10}$$

with the change in notation indicating that item difficulties $\beta_{ijt} = \beta_{it}$ are characterized as being constant across students.

Third-Level Model. The third level specifies the school-level of the model. At this level, estimates are provided of school-level ability estimates:

$$\gamma_{0t} = \gamma_{00} + \nu_t, \tag{2.11}$$

where γ_{00} is the average ability for school t, and ν_t follows $N(0, \zeta)$, in which ζ is the variance of ability at the school level. Item difficulty estimates are given as

$$\beta_{it} = \beta_i \tag{2.12}$$

indicating that at this level of the model item difficulties, $\beta_{it} = \beta_i$, are characterized as constant across schools. One useful outcome of this formulation is that the variance of ability can be decomposed into two components, a student-level component and a school-level component.

Kamata's model is specified for binary items, but it is also possible to incorporate polytomous items into this same three-level structure. Maier (2002) presented this for a partial credit model (PCM). Maier's multilevel PCM model is presented below at level one in the multilevel model. The multilevel PCM is reformulated here to enable it to be used in the MMixIRTM being developed in the present study.

First-Level Model. The first level identifies the measurement model. The PCM is based on the adjacent logit, and is an extension of the Rasch model.

$$\log\left[\frac{P_{im|m-1,m}(\theta_{jt})}{1 - P_{im|m-1,m}(\theta_{jt})}\right] = -(\theta_{jt} - b_{imjt}) \tag{2.13}$$

with $j = 1, \ldots, J$ examinees, $t = 1, \ldots, T$ schools, for a test of $i = 1, \ldots, I$ items, θ_{jt} is the ability of examinee j in school t, $P_{im|m-1,m}(\theta_{jt})$ is defined as

$$\frac{P_{im}(\theta_{jt})}{[P_{im-1}(\theta_{jt}) + P_{im}(\theta_{jt})]}, \tag{2.14}$$

where $P_{im}(\theta_{jt})$ is the probability of person j in a school t selecting category m of the item i correctly, and b_{imjt} is the item i step parameter. b_{imjt} has both j and t subscripts to enable its representation as regression coefficients in the HLM, even though the item step parameter is from the item characteristic curve (i.e., from the second level of the model). In the model (below), j and t subscripts are deleted at the second- and third-levels.

Second-Level Model. The second level, the student-level, provides student ability estimates,

$$\theta_{jt} = \gamma_{0t} + u_{jt}, \tag{2.15}$$

where γ_{0t} is the average ability of individual j in a school t, u_{jt} follows $N(0, \tau)$, in which τ is the variance of ability at the student level, and

$$b_{imjt} = b_{imt} \ . \tag{2.16}$$

This notation indicates that the item step parameters, $b_{imjt} = b_{imt}$, are constant across students.

Third-Level Model. Level 3 is the school-level of the model and provides estimates of school-level ability as

$$\gamma_{0t} = \gamma_{00} + \nu_t, \tag{2.17}$$

where γ_{00} is the average ability for school t, ν_t follows $N(0, \zeta)$, in which ζ is the variance of ability at the school level, and

$$b_{imt} = b_{im} \ . \tag{2.18}$$

At this level, the notation above indicates that the item step parameters, $b_{imt} = b_{im}$, are constant across schools.

The b_{im} in the multilevel PCM are the item step parameters. These can be decomposed into an overall item difficulty, δ_i, for item i. This is equivalent to the average of the xth step parameters for item i, and a category-specific mean deviation, η_{im}. That is,

$$b_{im} = \delta_i + \eta_{im}, \tag{2.19}$$

where

$$\delta_i = \frac{\sum_{v=1}^{x} b_{iv}}{x} \tag{2.20}$$

indicates the location of the item on the ability scale, and $\eta_{im} = b_{im} - \delta_i$ is the location of step x relative to the location of item i.

In this formulation, as in the previous one by Kamata, the variance of ability can be decomposed into two components, a student-level component and a school-level component.

Limitations of Multilevel IRT

In this section, some limitations of multilevel IRT models are discussed to provide motivation for the development of the MMixIRTM. One important limitation of multilevel IRT models is that they do not provide information about group membership beyond that provided by manifest predictors included in the model. As is noted below, this is consistent with previous approaches to detection of DIF. It is not necessarily the most useful way to detect DIF, however, if one is concerned with developing meaningfully related latent classes with respect to the causes of that DIF.

Information about comparison groups is a necessary feature of any DIF detection method and a central feature of the method to be developed in this study. We address the description of comparison groups in the context of a DIF analysis by using an approach that identifies latent groups in the data. Such groups are not immediately observable but share certain homogeneous response propensities that can be used to help explain how one latent group differs from another in item performance. It is these differences in response propensities that help to explain the causes of DIF.

Individuals may share a common set of response strategies, for example, when answering test questions (Mislevy & Verhelst, 1990; Rost, 1990). Likewise, schools within the same latent class may be homogeneous on certain criteria, while differing from schools in other school-level latent classes in particular and potentially important ways (Vermunt & Magidson, 2005). The incorporation of mixture models into multilevel IRT enables us to provide item and ability information for individuals in each student-level latent class as well as information about performance at the school-level. In this study, therefore, the multilevel IRT model is extended to include student-level and school-level mixtures. In addition, covariates are incorporated in the model at both the student-level and school-level to help describe characteristics of membership in the latent classes that are detected at each level.

2.5 Finite Mixture Models and Mixture IRT Models

2.5.1 Finite Mixture Models

Finite mixture models provide an appealing semi-parametric framework in which to model some unknown distributional shapes and model-based clustering (McLachlan & Peel, 2000). LCM is finite mixture modeling with categorical response variables. The terms finite mixture models and LCM are used interchangeably in this study. This kind of modeling began to emerge in 1960s in the sociology literature as a way of explaining respondent heterogeneity in survey response patterns involving dichotomous items. In LCM, latent classes are unobservable subgroups or segments in the data. Cases (e.g., individuals or students) within the same latent class are homogeneous on some relevant criteria, and cases in other latent classes are dissimilar on the same criteria.

The LCM can be parameterized in several ways. Heinen (1996) summarized three different parameterizations: conditional probability formulation, log-linear formulation, and log-linear with conditional probability formulation. These three LCM formulations are referred to in this study as unrestricted LCMs. Each formulation imposes restrictions on the model parameters in particular ways. By imposing some restrictions on a LCM, we have a restricted LCM. A LCM using log-linear parameterization with linear restrictions and an IRT model based on the cumulative logistic distribution provide comparable results (Heinen, 1996). In fact, however, there is still a difference between the two models in that the latent variable is discrete in the linear restricted LC model but continuous in the IRT model.

The log-linear formulation of the unrestricted LCM for the adjacent-categories logits is as follows:

$$\log\left[\frac{P_{im}(\theta_g)}{P_{i,m-1}(\theta_g)}\right] = (u_{im} - u_{i,m-1}) + u_{i,m-1\theta_g} \tag{2.21}$$

where $\frac{P_{im}(\theta_g)}{P_{i,m-1}(\theta_g)}$ is the logit in which the probabilities for responding in categories m and $m-1$ to item i, u_{im} and $u_{i,m-1}$, respectively, are category effects of item i, and $u_{i,m-1\theta_g}$ describes the difference between item i and the latent variable θ_g. If it can be assumed that

latent classes are ordered along a latent continuum, the G different logits (where G is the maximum number of latent classes) can be hypothesized to be linearly related to the latent variable (Heinen, 1996). This assumption can be formalized by imposing the following linear restriction on the $u_{i,m-1\theta_g}$:

$$u_{i,m-1\theta_g} = {u_{i,m-1}}^* \cdot \theta_g. \tag{2.22}$$

2.5.2 Mixture Item Response Theory Model

The LCM is based on the assumption of conditional independence of items, which is the same as the local independence assumption in IRT. The two models differ, however, in important ways. One important difference is that the latent ability in a LCM is assumed to be categorical whereas in IRT, it is assumed to be continuous. The result is that the MixIRTM, which is itself a combination of LCM and IRT, permits within-class variation on the latent variable θ. Members within a latent class formed using a MixIRTM, in other words, experience the same propensity for a response to each of the items on the test. That is, the IRT model within a latent class is the same for all members of the class, although examinees in the same class can vary on the underlying ability. In the LCM, however, examinees within a latent class are homogeneous in both their responses to items and on the latent ability.

Another difference between the two models is that there is an explicit functional relationship between the response probabilities and the latent variable in IRT while there is no such explicit functional relationship in LCM (Masters, 1985). The LCM also handles increase in test length differently than a MixIRTM. When test length increases, the matrix of response patterns becomes increasingly sparse. This sparseness can be a problem for LCMs because model parameters may not be well-estimated. The IRT portion in a MixIRTM potentially handles this sparseness somewhat better in that it imposes a parametric model (e.g., a cumulative logistic or normal ogive function) with strong assumptions describing the relationship between the response probabilities and the latent variable. The MixIRTM is also typically

seen to fit data better than conventional models of LCM or IRT (Muthén & Asparouhov, 2006).

Rost (1990, 1997) described a MRM in which an examinee population is assumed to be composed of a fixed number of discrete latent classes. In each latent class, a Rasch model is assumed to hold, but each class has different item difficulty parameters. In a MRM, members of the same latent class are assumed to experience the same relative difficulty among the items on the test, although within the same class members do differ in ability. The MRM also associates a class membership parameter, g, with each examinee that determines the relative difficulty of the items for that examinee, as well as a latent ability parameter, θ_{jg}, that influences the number of correct answers the examinee is expected to make to items on the test. The probability of a correct response in the MRM is written as

$$P(y_{ijg} = 1|g, \theta_{jg}) = \frac{1}{1 + \exp[-(\theta_{jg} - \beta_{ig})]} \ , \tag{2.23}$$

where g is an index for the latent class, $g = 1, \ldots, G$, $j = 1, \ldots, J$ examinees, θ_{jg} is the latent ability of a examinee j within class g, and β_{ig} is the Rasch difficulty parameter of item i for class g. Rost (1990) suggested that the primary diagnostic potential of the MRM is in its use for accounting for qualitative differences among examinees, and its simultaneous ability to quantify that ability with respect to the same items.

von Davier and Yamamoto (2004) extended Masters's PCM and Muraki's generalized partial credit model (GPCM) to the class of discrete mixture models. The mixture PCM, a Rasch model, is formulated as follows:

$$P_P(y_{ijg} = x|g, \theta_{jg}) = \prod_{m=1}^{M_i} \frac{\exp[\sum_{v=0}^{x}(\theta_{jtg} - b_{img})]}{\sum_{h=0}^{M_i} \exp[\sum_{v=0}^{h}(\theta_{jtg} - b_{img})]}, \tag{2.24}$$

where $g = 1, \ldots, G$ student-level latent classes, and $j = 1, \ldots, J$ examinees. The polytomous item scores are $0, 1, \ldots, M_i$ for the $i = 1, \ldots, I$ items. x is the given item score. θ_{jtg} is the ability of examinee j in school t in the class g, and b_{img} is the item step parameter of the category m of item i for the class g.

Extending Mixture IRT Models. An important limitation of MixIRTMs is that they essentially ignore the basic multilevel structure that is present beyond the student-level in much of educational test data. In this study, we incorporate that multilevel structure into a MixIRTM, and extend the model into a multilevel MixIRTM (MMixIRTM). We then use this model for DIF detection at both the student- and school-levels, focusing on detection of latent classes of students and school whose performance on particular items differs along lines that are not necessarily associated with membership in a particular manifest group. The approach proposed in this study provides DIF information at the student- and school-levels along with information describing the composition of the different groups in the DIF comparisons. In this way, information can be provided that will enable school personnel or others involved in interpretation of test results to easily describe the composition of comparison groups, including other schools like themselves.

2.6 Multilevel Latent Class Models

An alternative approach to DIF detection is found in LCMs (Webb, Cohen, & Schwanenfluegel, in press). The rationale for use of a LCM is the same as that proposed by Cohen and Bolt (2005). The difference is that a LCM is used in place of an IRT model. Webb et al. demonstrated that nuisance dimensionality was responsible for what was previously perceived as a bias on the Peabody Picture Vocabulary Test III (Dunn & Dunn, 1997) against African-American preschool children. Using a LCM, Webb et al. showed that the actual cause of the DIF was due to use of particular response strategies by latent classes of examinees.

An important limitation of the LCM, that is shared with IRT models, is it assumes that observations are independent. Unfortunately, this assumption is often violated. Multilevel LCMs have been suggested that relax this assumption (Vermunt, 2003; Vermunt & Magidson, 2005; Bijmount, Paas, & Vermunt, 2004; Asparouhov & Muthén, 2007). This type of model uses a discrete unspecified distribution, which provides a nonparametric multilevel LCM.

In the context of the PSAT/NMSQT, this type of model would allow for both student-level latent classes and school-level latent classes sharing the same parameter values. This approach imposes less restrictive distributional assumptions and typically provides faster and more stable estimation than non-linear regression models (Vermunt, 2003).

As Vermunt (2003) notes, nonparametric does not mean distribution free. Rather, nonparametric is used to indicate that the normal distribution assumption is replaced by a multinomial distribution assumption. The MLCM can be described as follows:

$$P(y_{ijt}) = \sum_{k=1}^{K} \sum_{g=1}^{G} \pi_k \cdot \pi_{g|k} \cdot P(y_{ijtgk}|\theta_{gk}), \tag{2.25}$$

where $g = 1, \ldots, G$ student-level latent classes, $k = 1, \ldots, K$ school-level latent classes, $\pi_{g|k}$ indicates the relative sizes of latent classes at the student-level conditional on latent class membership at the school level, π_k is the proportion of schools for each class, $P(y_{ijtgk})$ is the conditional probability to item i for latent classes g and k.

$P(y_{ijtgk}|\theta_{gk})$ in a LCM can be restricted or unrestricted. An unrestricted LCM may not always be the most useful, however, when the analysis is intended to be for measurement rather than for data reduction (Heinen, 1996; Masters, 1985). A number of restricted LCMs can be obtained to investigate the relationship between the latent variable and manifest indicators such as items.

As mentioned earlier, Heinen (1996) reported equivalence of restricted LCMs and IRT models using a log-linear parametrization for the LC model and a cumulative logistic function for the IRT model. Vermunt's modeling using a restricted LCM as the measurement model (Vermunt, 2003; Bijmount, Paas, & Vermunt, 2004) is similar to the multilevel IRT model. The MixIRTM portion of a MMixIRTM, however, can be more easily viewed as a combination of an unrestricted and a restricted LCM. This combined model combines a discrete LCM with a continuous IRT model.

Comparison of Kamata's multilevel IRT model with a restricted multilevel latent class model. The comparison of Kamata's (2001) multilevel IRT model with a multilevel restricted LCM is revisited here to illustrate the meaning of nonparametric as

used in this study. The ability structure of a multilevel IRT model can be given as follows:

$$\theta_{jt} = \gamma_{00} + u_{jt} + \nu_t, \tag{2.26}$$

where $j = 1, \ldots, J$ students, $t = 1, \ldots, T$ schools, γ_{00} is the average ability for school t, u_{jt} is the ability across students, u_{jt} follows $N(0, \tau)$, in which τ is the variance of ability at the student level, and ν_t follows $N(0, \zeta)$, in which ζ is the variance of ability at the school level. The total variance of ability in the multilevel IRT model is decomposed into the student-level ability variance (i.e., τ) and the school-level ability variance (i.e., ζ).

The ability structure of a MLCM can be given as follows:

$$\theta_{jt} = \sum_{k=1}^{K} \sum_{g=1}^{G} \pi_k \cdot \pi_{g|k} \cdot \theta_{gk}, \tag{2.27}$$

where $g = 1, \ldots, G$ student-level latent classes, and $k = 1, \ldots, K$ school-level latent classes.

The multilevel restricted LCM can be considered a nonparametric version of a multilevel IRT model in that class-specific ability, θ_{gk} (where g indicates latent class at level 2 and k indicates latent class at level 3), in the multilevel restricted LCM is similar to the random effect of ability, θ_{jt} in the IRT model. The gk subscript on θ for the LCM indicates class specific ability whereas the jt subscript in Equation (2.26) is a random effect specific to student j in school t.

This perspective is of interest in that a multilevel LCM can be considered as a nonparametric approach to the multilevel IRT portion of a MMixIRTM. We employ the point of view in this study, however, that the MMixIRTM is an extension of a multilevel unrestricted LCM by including an IRT model.

Limitations of a Multilevel Latent Class Model. One concern with a multilevel LCM is there is no variation in ability across latent classes. That is, θ_{gk}, is the same for all examinees within a latent class in a MLCM. The probability for giving a response on a particular item i, conditional on membership in a latent class is the same for all students in that latent class. The proposed MMixIRTM allows for variability in ability within latent classes.

Asparouhov and Muthén (2007) extended the structural equation model and factor mixture model (which can be formulated as a MixIRTM) into a multilevel model. This approach and a MMixIRTM approach are similar in that both use a factor mixture model in the multilevel extension. In other words, both approaches allow examinees to vary on the latent ability. The group level mixture is a parametric random effect (i.e., normally distributed) rather than a nonparametric model in Asparouhov and Muthén (2007).

2.7 Multilevel mixture item response theory model

2.7.1 Model rationale

The proposed MMixIRTM estimates mixtures of latent classes at two levels, a student-level and a school-level. Student-level latent classes capture the association between the responses at the student-level unit. The MixIRTM assumes that there may be heterogeneity in response patterns at the student-level which should not be ignored (Mislevy & Verhelst, 1990; Rost, 1990). The MMixIRTM does not exclude the possibility, however, that there may be no student-level latent classes. It is interesting to note that, if no student-level latent classes exist, this would indicate that there would also be no school-level latent classes. The reason is that school-level units are clustered based on the likelihood of their students belonging to one of the latent classes. In a MMixIRTM as presented in this study, in other words, it is not meaningful to have school-level classes if no student-level latent classes are present.

Viewed in this way, school-level latent classes capture the association between the students within school-level units. Latent classes at the school-level, however, may differ in the probability that students belong to particular latent classes. This is accommodated in the MMixIRTM by allowing for the possibility that school-level latent classes may differ in the proportions of students in each student-level latent class contained in a school-level latent class.

The modeling of ability variance in the MMixIRTM is different from that of a multilevel IRT model in that the total variance of ability is modeled rather than the component

variances. A special case of the MMixIRTM may arise if there are the same proportions of student-level latent classes in each school-level latent class and there is only student-level class-specific item difficulties. In this case, the total ability variance can be decomposed into student-level and school-level components. This is not usually the case, however. A more realistic possibility is that the proportions of student-level latent classes will be different in each school-level latent class. In this case, the total variance of ability can not be decomposed. The hierarchical data structure of ability variance allows the estimation of both student- and school-specific variance of ability. This approach to ability specification provides not only class-specific information, but also can be used to approximate a continuous mixture distribution (Vermunt, 2003). If there are large differences between schools with respect to the proportion of student-level latent classes, it indicates that the students belonging to the same school-level latent class are strongly homogeneous (Vermunt & Magidson, 2005).

Ability Structure of MMixIRTM. Equation (2.7) is revisited here to describe the structure of data in terms of students, nested within school with probabilities of mixtures at the student- and school-level literally and graphically.

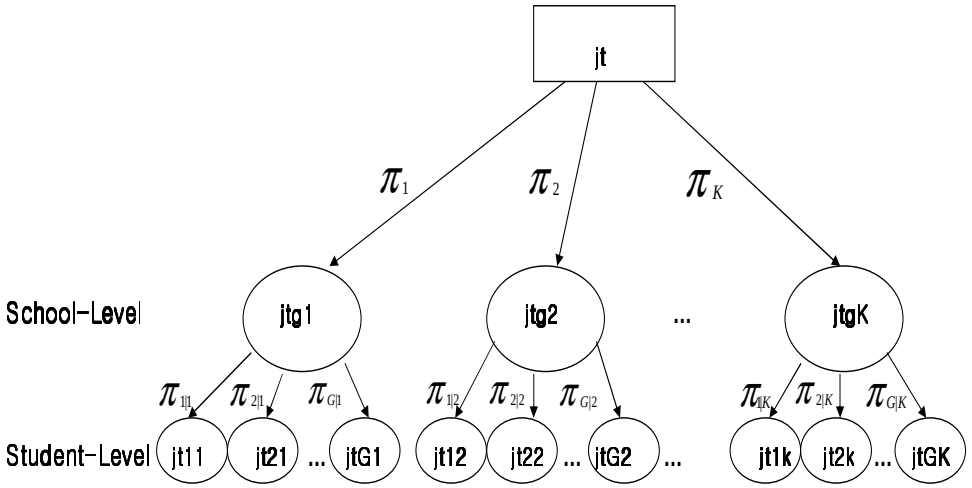
$$\theta_{jtgk} \sim N(\mu_{gk}, \sigma^2_{gk}) \ . \tag{2.28}$$

Each student has manifest identification, j and t. Student j is nested within school t. In addition, he/she has latent identification, g and k. Abilities of students (i.e., θ_{jtgk}) have variability for each mixture unit of g and k, which has its own mean (i.e., μ_{gk}) and variability (i.e., σ^2_{gk}). Students are composed with student-level probability of mixture (i.e., $\pi_{g|k}$) and school-level probability of mixture (i.e., π_k). This structure is represented in Figure 2.3.

Proportion Structure of MMixIRTM. The structure of $\pi_{1:G|k}$ is shown in Table 2.2. There are K probability arrays $\pi_{1:G|k}$, $k = 1, \cdots, K$, where G is the dimension of each array.

Item Difficulty Structure of MMixIRTM. In the general model of a MMixIRTM, item difficulty parameters have both student- and school-level class-specific values. These can be represented as β_{igk}. A graphical visualization of this relationship is shown in the three-dimensional array in Figure 2.4. The column in Figure 2.4 represents items (i.e., i),

Figure 2.3: Graphical Representation of Data Structure



the row represents the student-level mixture (i.e., g), and the school-level mixture (i.e., k), is represented by the dotted line. Figure 2.4 is intended to help visualize the item difficulty structure, so it can be seen that student-level class-specific item difficulties stack up at the school-level mixtures.

Probability Structure of MMixIRTM. The probability of getting a correct response in the MMixIRTM can be given as follows:

$$P_{ijt} = \sum_{k=1}^{K} \sum_{g=1}^{G} \pi_k \cdot \pi_{g|k} \cdot P(y_{ijtgk} = 1 | g, k, \theta_{jtgk}). \tag{2.29}$$

In the structure shown in Figure 2.5, the circles represent the latent variable (i.e., ability) and the rectangles represent the manifest variable (i.e., item responses). The arrows from g and k indicate that the probability of getting a correct response is a function of g and k.

Table 2.2: Proportion Structure of MMixIRTM

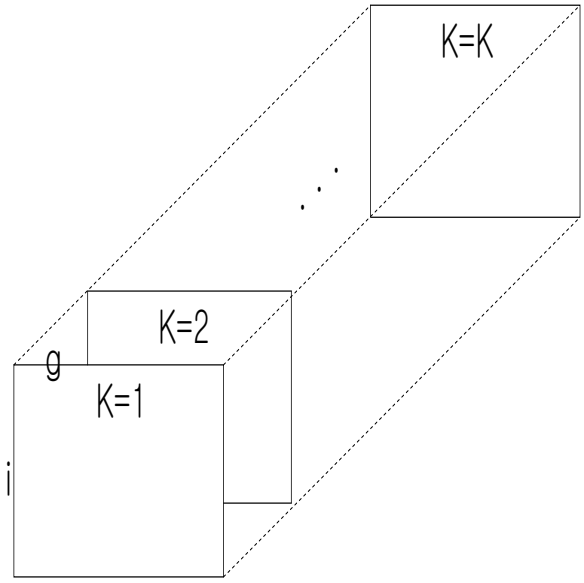
	$K=1$	$K=2$	. .	$K=K$
$G=1$	$\pi_{1\|1}$	$\pi_{1\|2}$	. .	$\pi_{1\|K}$
$G=2$	$\pi_{2\|1}$	$\pi_{2\|2}$	. .	$\pi_{2\|K}$
.	.	.	. .	.
.	.	.	. .	.
$G=G$	$\pi_{G\|1}$	$\pi_{G\|2}$	. .	$\pi_{G\|K}$
Sum	$\sum_{g=1}^{G}\pi_{g\|1}=1$	$\sum_{g=1}^{G}\pi_{g\|2}=1$	. .	$\sum_{g=1}^{G}\pi_{g\|K}=1$

Meaning of the term Multilevel in MMixIRTM. In this subsection, the term multilevel is explained to emphasize the model rationale. This term can be used with both mixture and IRT. As described earlier for the model rationale, the school-level mixture is composed of different proportions of student-level mixtures. The result is that ability and item difficulty are modelled with both student- and school-levels mixtures. In this case, the term, multilevel is used for the description of the two-level for mixture. If there are the same proportions of student-level latent classes in each school-level latent class as a special case of the MMixIRTM, the term multilevel is used for the description of multilevel IRT model.

The three sections of the PSAT/NSMQT, the critical reading, mathematics, and writing skills sections, include multiple-choice, and gridded-response items. The MMixIRTM was used in this study only for the binary response items. The model formulation for each multiple-choice, short and extended response item, however, is described for the example.

The MMixIRTM developed in this study is described below from three different perspectives to show how the MMixIRTM can be formulated by adding the flaw feature to the combined models like a MixIRTM, a multilevel IRT model, and a MLCM. First, a MMixIRTM is formed as the incorporation of mixtures into a multilevel IRT model. Second, a MMixIRTM

Figure 2.4: Visualization of Class-Specific Item Difficulties



is formed by incorporating multilevel structure into a MixIRTM. Finally, a MMixIRTM is formed of an IRT model incorporated into a multilevel unrestricted LCM.

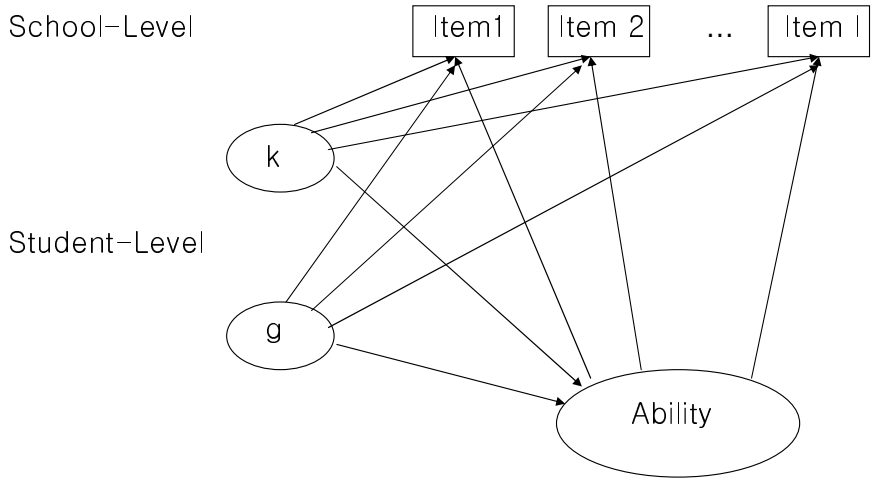
2.7.2 The Multilevel Mixture IRT Model as an Extension of a Multilevel IRT Model

The following MMixIRTM formulation is based on the multilevel IRT model described in the previous section. The first part presents the MMixIRTM for binary items.

First-Level Model. At the first level of the model, the measurement model is identified:

$$\log\left[\frac{P_i(\theta_{jtgk})}{1-P_i(\theta_{jtgk})}\right] = \theta_{jtgk} - \beta_{ijtgk} \tag{2.30}$$

Figure 2.5: MMixIRTM Diagram



where $g = 1, \ldots, G$ student-level latent classes, $k = 1, \ldots, K$ school-level latent classes, $j = 1, \ldots, J$ examinees, and $t = 1, \ldots, T$ schools, for a test of $i = 1, \ldots, I$ items, θ_{jtgk} is the class-specific ability of examinee j in school t, $P_i(\theta_{jtgk})$ is the probability of person j in a school t answering the item i correctly, and β_{ijtgk} is the class-specific difficulty of item i. As was the case for Kamata's (2001) model (described earlier), β_{ijtgk} at the first level has both j and t subscripts, even though item difficulty is an item characteristic and is not actually estimated at this level. This form of representation for item difficulty is due to the regression

coefficients present for each component in the model at the first level in the HLM. The j and t subscripts are deleted at the second- and third-level respectively.

Second-Level Model. The second level is the student-level model. Student ability estimates are given as follows:

$$\theta_{jtgk} = \gamma_{0tgk} + u_{jtgk}, \tag{2.31}$$

where γ_{0tgk} is the class-specific average ability of individual j in school t, u_{jtgk} is distributed $N(0, \tau_{g.})$, where $\tau_{g.}$ is the variance of ability at the student -level, and

$$\beta_{ijtgk} = \beta_{itgk} \tag{2.32}$$

are class-specific difficulties $\beta_{ijtgk} = \beta_{itgk}$. As indicated in the notation, these are characterized at the second level as being constant across students.

Third-Level Model. The third level specifies the school-level model and provides estimates of school-level ability estimates:

$$\gamma_{0tgk} = \gamma_{00gk} + \nu_{tgk}, \tag{2.33}$$

where γ_{00gk} is the class-specific average ability, ν_{tgk} follows $N(0, \zeta_{.k})$, which $\zeta_{.k}$ is the variance of ability at the school level, and

$$\beta_{itgk} = \beta_{igk} \tag{2.34}$$

are item difficulties $\beta_{itgk} = \beta_{igk}$ which are characterized at this level as constant across schools.

As was noted above for the multilevel IRT model, polytomous items also can be incorporated into the MMixIRTM. This is done as follows:

First-Level Model. The first level identifies the measurement model. The probability of a response to category m for item i is defined as a function of ability, θ, and an item step parameter, b:

$$\log\left[\frac{P_{im|m-1,m}(\theta_{itgk})}{1 - P_{im|m-1,m}(\theta_{itgk})}\right] = -(\theta_{jtgk} - b_{imjtgk}) \tag{2.35}$$

where $g = 1, \ldots, G$ student-level latent classes, $k = 1, \ldots, K$ school-level latent classes, $j = 1, \ldots, J$ examinees, and $t = 1, \ldots, T$ schools. The polytomous item scores are $0, 1, \ldots, M_i$ for the $i = 1, \ldots, I$ items. θ_{jtgk} is the ability of examinee j in school t and in both latent classes g and k, $P_{im|m-1,m}(\theta_{jtgk})$ is defined as

$$\frac{P_{im}(\theta_{jtgk})}{[P_{i,m-1}(\theta_{jtgk}) + P_{im}(\theta_{jtgk})]}, \tag{2.36}$$

where P_{im} is the probability of person j in a school t answering the category m of the item i correctly, b_{imgk} is the difficulty of the category m item i for classes g and k.

Second-Level Model. The second level, the student-level model, provides student ability estimates:

$$\theta_{jtgk} = \gamma_{0tgk} + u_{jtgk}, \tag{2.37}$$

where γ_{0tgk} is the average ability of individual j in a school t for both latent classes g and k, u_{jtgk} follows $N(0, \tau_{g.})$, which $\tau_{g.}$ is the variance of ability at the student level, and

$$b_{imjtgk} = b_{imtgk} \tag{2.38}$$

indicating that item difficulties $b_{imjgkt} = b_{imtgk}$ are treated as constant across students in both classes g and k.

Third-Level Model. Level 3 specifies the school-level model and provides estimates of school level ability estimates:

$$\gamma_{0tgk} = \gamma_{00gk} + \nu_{tgk}, \tag{2.39}$$

where γ_{00gk} is the average ability for school t in both classes g and k, ν_{tgk} follows $N(0, \zeta_{.k})$, which $\zeta_{.k}$ is the variance of ability at the school level,

$$b_{imtgk} = b_{imgk} \tag{2.40}$$

and item difficulties $b_{imtgk} = b_{imgk}$ are characterized as constant across schools in latent classes.

The b_{imgk} in the multilevel mixture PCM are the item step parameters. These can be decomposed into an overall item difficulty, δ_{igk}, for item i in latent group g and latent group

k, which is equal to the average of the xth step parameters for item i, and a category-specific mean deviation, η_{imgk}. That is,

$$b_{imgk} = \delta_{igk} + \eta_{imgk} \ , \tag{2.41}$$

where

$$\delta_{igk} = \frac{\sum_{v=1}^{x} b_{ivgk}}{x} \tag{2.42}$$

is the item's location on the ability scale, and $\eta_{imgk} = b_{imgk} - \delta_{igk}$ is the location of step x relative to the location of item i.

In the multilevel IRT formulation of Kamata (2001), the variance of ability is decomposed into two components, a student-level component, τ, and a school-level component ζ. Even though the formulation above in Equations (2.37) and (2.39) is based on the same multilevel IRT model, the modeling of the variance of ability in the MMixIRTM is different in that the total ability variance is not decomposed into student-level and school-level ability variances. If the proportions in student-level latent classes are the same across school-level latent classes, then the total ability variance can be decomposed into student-level and school-level variances. If this is not the case, however, then the total ability variance can not be decomposed, because the proportion of student-level latent classes is different across school-level latent classes. Since this is the likely case with real data, the structure of θ_{jtgk} is as follows:

$$\theta_{jtgk} \sim N(\gamma_{00gk}, \sigma^2_{gk}), \tag{2.43}$$

where γ_{00gk} is the mean of ability for classes g and k, and σ^2_{gk} is the variance of ability for classes g and k, which sum $\tau_{g.}$ and $\zeta_{.k}$. γ_{00gk} in Equation (2.43) is the same as μ_{gk} in Equation (2.49).

2.7.3 A Multilevel Mixture IRT Model Extension of a Mixture IRT Model

Equation (2.44) presents a MixIRTM with mixtures at the student-level and school-levels. The probability of a correct response is given as

$$P_B(y_{ijtgk} = 1|g, k, \theta_{jtgk}) = \frac{1}{1 + \exp[-(\theta_{jtgk} - \beta_{igk})]}, \tag{2.44}$$

where $g = 1, \ldots, G$ student-level latent classes, $k = 1, \ldots, K$ school-level latent classes, $j = 1, \ldots, J$ examinees, and $t = 1, \ldots, T$ schools, for a test of $i = 1, \ldots, I$ items. θ_{jtgk} is the ability of examinee j in school t and in both latent classes g and k, and β_{igk} is the difficulty of item i for classes g and k.

Below, we describe a multilevel mixture PCM for constructed-response items. We modify this formulation in the sequel depending on the types of item response formats. The model formulation is as follows:

$$P_P(y_{ijtgk} = x|g, k, \theta_{jtgk}) = \prod_{m=1}^{M_i} \frac{\exp[\sum_{v=0}^{x}(\theta_{jtgk} - b_{ivgk})]}{\sum_{h=0}^{M_i} \exp[\sum_{v=0}^{h}(\theta_{jtgk} - b_{ivgk})]}, \tag{2.45}$$

where $g = 1, \ldots, G$ student-level latent classes, $k = 1, \ldots, K$ school-level latent classes, $j = 1, \ldots, J$ examinees, and $t = 1, \ldots, T$ schools. The polytomous item scores are $0, 1, \ldots, M_i$ for the $i = 1, \ldots, I$ items, x is the item score, θ_{jtgk} is the ability of examinee j in school t and in both latent classes g and k, and b_{ivgk} is the difficulty of the category m item i for classes g and k.

The b_{ivgk} in the multilevel mixture PCM are the item step parameters. These can be decomposed into an overall item difficulty, δ_{igk}, for item i in latent groups g and k, which is equal to the average of the xth step parameters for item i, and a category-specific mean deviation, η_{ivgk}. That is,

$$b_{ivgk} = \delta_{igk} + \eta_{ivgk}, \tag{2.46}$$

where

$$\delta_{igk} = \frac{\sum_{v=1}^{x} b_{ivgk}}{x} \tag{2.47}$$

is the item's location on the ability scale and $\eta_{ivgk} = b_{ivgk} - \delta_{igk}$ is the location of step x relative to the location of item i.

For both binary and polytomous items, θ_{jtgk} has the following structure:

$$\theta_{jtgk} \sim N(\mu_{gk}, \sigma^2_{gk}) \tag{2.48}$$

where μ_{gk} is the mean of ability for classes g and k, and σ^2_{gk} is the variance of ability for classes g and k.

2.7.4 A Multilevel Mixture IRT Model as an Extension of a Multilevel LCM

The following presents a multilevel unrestricted LCM with an IRT model extension. This combination provides a MixIRTM in the MMixIRTM. It leads to the same model formulation as the previous two perspectives, i.e., the incorporation of mixtures into a multilevel IRT model and the incorporation of a multilevel structure into a MixIRTM. It is of interest to compare the θ structure of MMixIRTM among these three formulations.

As we can see below, θ_{jtgk} follows $N(\mu_{gk}, \sigma^2_{gk})$, indicating that there is variability in θ within classes g and k. This θ structure is different from the MMixIRTM in Vermunt's (2003) multilevel LCM. In Equation (2.25) of Vermunt's MLCM, there is no variability in ability, θ_{gk}. However, in the MMixIRTM, there is variability in ability, $N(\mu_{gk}, \sigma^2_{gk})$, as is shown below:

$$\theta_{jtgk} \sim N(\mu_{gk}, \sigma^2_{gk}) \tag{2.49}$$

where $j = 1, \ldots, J$ students, $t = 1, \ldots, T$ schools, $g = 1, \ldots, G$ student-level latent classes, $k = 1, \ldots, K$ school-level latent classes, μ_{gk} is the mean of ability for classes g and k, and σ^2_{gk} is the variance of ability for classes g and k.

2.7.5 A Multilevel Mixture IRT Model with Covariates

In this section, we extend the MMixIRTM to a finite mixture regression model that incorporates covariates. Covariates can be particulary useful for helping to identify latent classes in the data (Smit, Kelderman, & van der Flier, 1999). Smit et al. also indicated that item parameters and group memberships were recovered better as the strength of the association between the latent class and the covariates increased.

The extension described below is a finite mixture regression model that includes covariates. As indicated above, these covariates can be used both to predict latent class membership for individual j and school t and to describe the relationship between the means for individual j on each of the G latent class means and the relationship between the means for school t on each of the k latent class means.

Using covariates in the MMixIRTM provides a change in the usual approach for detecting membership in latent classes. This latter approach follows a two-step procedure in which a class is estimated for each examinee in Step 1 and then in Step 2, cross-tabulation, regression, or some other method is used to relate membership in a latent group to one or more covariates. One disadvantage of this two-step approach is that it permits estimation errors to intrude on the classification process which attenuate the relationship between the covariate(s) and the latent classes (Vermunt & Magidson, 2005). Smit et al. (1999) demonstrate how standard errors can be reduced and the accuracy of classification improved by incorporating variables associated with the latent class variable into the model. Cho, Cohen, and Kim (2006) found that the use of a covariate improved the recovery of both item difficulties and group membership, when there was an overlap of 80 percent or more among the covariate and the latent groups, and when no guessing was simulated. Including such covariates in the same model, in other words, helped to reduce the attenuation that can occur in the two-step approach mentioned above.

The MMixIRTM with covariates is modeled as followed: Students and schools are modeled as belonging to the latent class for which they have the highest probability of membership. The proportion of individuals in class g conditional on school level latent class, $\pi_{g|k}$, is

$$\pi_{g|k,W_j} = \frac{\exp(\gamma_{0gk} + \sum_{p=1}^{P} \gamma_{pg} W_{jp})}{\sum_{g=1}^{G} \exp(\gamma_{0gk} + \sum_{p=1}^{P} \gamma_{pg} W_{jp})}, \tag{2.50}$$

where $\pi_{g|k}$ is the probability of an individual belonging to a class g given k. The G latent groups are modeled as functions of the individual-level covariates W_j, and $\pi_{g|k}$ is a multinomial logistic regression. γ_{pg} is a class-specific effect of covariate p on group membership. For identifiability, $\gamma_{01k} = 0$ for all ks and $\gamma_{p1} = 0$.

In addition, the probability of a school belonging to latent class k, π_k, can be written as

$$\pi_{k|W_t} = \frac{\exp(\gamma_{0k} + \sum_{p=1}^{P} \gamma_{pk} W_{tp})}{\sum_{k=1}^{K} \exp(\gamma_{0k} + \sum_{p=1}^{P} \gamma_{pk} W_{tp})}, \tag{2.51}$$

where K latent groups are modeled as functions of the school-level covariates, W_t, and γ_{pk} is a class-specific effect of covariate p on group membership. For identifiability, $\gamma_{01} = 0$ and $\gamma_{.(k=1)} = 0$.

2.7.6 Special Cases of the Multilevel Mixture IRT Model with Covariates

Below, we introduce three special cases of the MMixIRTM which have some utility for estimation of DIF in the multilevel model. These are each obtained through the use of different sets of constraints.

Special Case I. The first special case of a MMixIRTM is that in which the proportions of student-level latent classes are same among school-level latent classes. When this assumption holds, item (i.e., β_{igk} and b_{igk}) and ability (i.e., θ_{jtgk}) parameters can be split into student-level and school-level parameters. We illustrate this special case as follows for binary items:

$$P_B(y_{ijtgk} = 1|g, k, \theta_{jtg}, \theta_{jtk}) = \frac{1}{1 + \exp[-(\theta_{jtg} + \theta_{jtk}) - (\beta_{ig} + \beta_{ik})]} \ . \tag{2.52}$$

Equation (2.52) shows that the item difficulty parameters β_{ig} and β_{ik} are estimated separately for the student-level and for the school-level latent groups. If this assumption holds, this formulation indicates that DIF can be analyzed separately at the student-level and at the school-level.

Special Case II. The second special case we consider is that for which item and ability parameters do not vary across school-level classes. This model can be useful when the purpose of analysis is on identifying different students' strategies with incorporating multilevel data structure. We illustrate this special case as follows for binary items:

$$P_B(y_{ijtgk} = 1|g, k, \theta_{jtg}, \theta_{jt}) = \frac{1}{1 + \exp[-(\theta_{jg} + \theta_{jt} - \beta_{ig})]} \ . \tag{2.53}$$

It can be seen in Equation (2.53) that the item difficulty parameter β_{ig} differs among student-level latent classes. It does not contain a k subscript indicating that the same estimates hold

for each school-level latent class. If this model holds, then DIF is tested only at the student-level. The j subscript in θ_{jt} of Equation (2.53) indicates students in school t of class g. A similar model was described in Asparouhov and Muthén (2007).

Special Case III. The third special case of the MMixIRTM is one in which item and ability parameters do not vary among student-level classes. This special model is of interest in case we seek to obtain school-level DIF information. In fact, this model contains information aggregated across student-level latent classes for each school-level latent class. We illustrate this special case for binary items as follows:

$$P_B(y_{ijtgk} = 1|g, k, \theta_{jt}, \theta_{jtk}) = \frac{1}{1 + \exp[-(\theta_{jt} + \theta_{jtk} - \beta_{ik})]}. \tag{2.54}$$

The item difficulty estimates, β_{ik}, contain a k subscript but not a g subscript indicating that they differ only by school-level latent class. This formulation was illustrated in Vermunt (2007), and is of interest when we seek to examine DIF only among the school-level latent classes.

The general MMixIRTM described earlier and the three special cases of the general model are summarized in Table 2.3.

In the next section, DIF will be defined with respect to the MMixIRTM as the general model and the three special cases, respectively.

2.7.7 DIF Analysis using Multilevel Mixture IRT Model

Kamata and Binici (2003) extended a two-level DIF model to a three-level DIF model using a hierarchical generalized linear model (HGLM) framework. The three-level model accounted for variation of DIF across schools as well as identification of school characteristics that help explain such variation. The magnitude of DIF was modeled across schools by Kamata et al. (2005) and included school characteristics (i.e., predictors) to determine whether they can help explain the variation of DIF among schools. Kamata et al. also modeled the magnitude of DIF at the school-level with random effects and described it with school-level predictors.

Table 2.3: Comparisons of the Multilevel Mixture IRT Model

Model	Model Rationale	Level	Ability Distribution	Proportions	Item Difficulty
General Model	k has different proportions of gs.	Both Student- and School-Levels	$\theta_{jtgk} \sim N(\mu_{gk}, \sigma^2_{gk})$	$\pi_{g\|k}, \pi_k$	β_{igk}
Special Case I	k has the same proportions of gs.	Student-Level	$\theta_{jtg} \sim N(\mu_g, \sigma^2_g)$	π_g	β_{ig}
		School-Level	$\theta_{jtk} \sim N(\mu_k, \sigma^2_k)$	π_k	β_{ik}
Special Case II	gs are clustered	Student-Level	$\theta_{jg} \sim N(\mu_g, \sigma^2_g)$	π_g	β_{ig}
(Muthén, & Asparouhov, 2007)	with respect to same student-level ability distribution.	School-Level	$\theta_{jt} \sim N(0, 1)$	NA	NA
Special Case III	ks are clustered	Student-Level	$\theta_{jt} \sim N(0, 1)$	NA	NA
(Vermunt, 2007)	with respect to same school-level ability distribution.	School-Level	$\theta_{jtk} \sim N(\mu_k, \sigma^2_k)$	π_k	β_{ik}

That is, Kamata et al. modeled a three-way interaction of item difficulty × student-level manifest group × school characteristics for the item difficulty parameters.

This approach to detection of school-level DIF is different from that of the proposed model in two ways: First, it is based on the use of manifest groups, and second, the modeling of DIF is incorporated directly into estimation of item difficulty. This approach by Kamata et al. is markedly different from the usual approach which detects DIF by calculating some function indicating a difference between item parameters among the different manifest groups.

The MMixIRTM proposed in this study functions differently from that of either the standard DIF detection approach or the one described by Kamata et al. (2005). Kamata et al. (2005) examined DIF based on item performance differences among manifest groups. In the current study, DIF is considered by examining differential item performance among different latent classes of students and schools. Performance in this approach is not directly associated with membership in a particular manifest group and group-specific item difficulties are present for both student-level and school-level latent classes.

Four Different Types of DIF. DIF can be investigated at both student-level and school-level for the general MMixIRTM as well as for each of the three special cases described above. For the general MMixIRTM, item difficulty and item step parameters have both subscripts g and k. This indicates that each school-level latent class, k, has its own set of student-level latent classes. This in turn indicates that item parameters need to be represented for each student-level latent class and for each school-level latent class.

In the general MMixIRTM, both student-level DIF and school-level DIF can be detected. School-level DIF is defined as the differences between the item difficulties given the student-level latent classes. Student-level DIF can be detected by calculating and testing the differences between the item difficulties (i.e., β_{igk}) within a school-level latent class. For polytomous items, item step parameters, b_{ivgk}, can be decomposed into a class-specific item difficulty, δ_{igk}, and a category-specific mean deviation, η_{imgk}. In the general MMixIRTM, it is assumed that item difficulties, δ_{igk}, may differ among latent classes. The category param-

eters, η_{imgk}, are common for latent classes which results in the formulation of η_{imgk} as η_{im}. This formulation is consistent with the approach in Muraki (1999) for DIF in the PCM among manifest groups. Thus, student-level DIF for polytomous items is defined as the difference between the overall item difficulties (i.e., δ_{igk}) within a school-level latent class. School-level DIF for polytomous items is defined as the difference between item difficulties for the school-level latent classes given the student-level latent classes.

Second, both student-level DIF and school-level DIF can be detected for Special Case I, but only if the proportions of student-level latent class are same across school-level latent classes. As noted above, student-level DIF is detected as differences among the item difficulties, β_{ig} and δ_{ig} for student-level latent classes, g. School-level DIF is detected from the differences among the item difficulties, β_{ik}, among school-level latent groups, k. One caveat with respect to Special Case I is that it is somewhat unrealistic to assume that the same proportions of student-level latent classes exist in each school-level latent class.

Third, as noted above for Special Case II, only student-level DIF is defined. As mentioned earlier, this approach, which can be useful if group-level information is not needed, is done by nonparametrically incorporating the hierarchical structure of the data into the model.

Fourth, for Special Case III, only school-level DIF is possible. As noted earlier, this DIF is detected as differences among the item difficulties, β_{ik} and δ_{ik}. This version of DIF is somewhat less informative than that of the general MMixIRTM as differences among student-level item difficulty estimates are aggregated within each school-level latent class. That is, β_{igk} in the general MMixIRTM is only available as $\beta_{i.k} = \beta_{ik}$ in Special Case III. In other words, information is not available on item characteristics meaning that student-level covariates, W_{jp}, do not enter into the model for this special case.

Although four different DIF analyses are available with the general MMixIRTM and the three special cases described above, DIF analysis using the MMixIRTM is of primary interest. This is because the group-level information for both student-level and group-level

DIF is of most interest and more flexibility is allowed in terms of student-level proportions. The specific DIF analysis procedures to be used will be presented in the next chapter.

Multiple-Group DIF Among Latent Classes. The current SOAS report provides performance comparisons between a given school, defined as the local group, and a reference group that is used to represent expected academic performance for the given school. The DIF analysis approach in this study, however, is different from the usual multiple-group DIF analysis in that there is a potential for more than one reference group, when there are more than three school-level latent classes (i.e., when $K \geq 3$). The number of reference groups is calculated in this study as $K - 1$. The given school is included in one school-level latent class and the class-specific item difficulty parameters, β_{igk} or b_{igk}, for the given school can be compared with those for the other latent classes (i.e., with the $\beta_{igk'}$ or $b_{igk'}$). In other words, since school characteristics are to be clustered with respect to particular performance characteristics, the given school is compared to schools with different performance characteristics. This approach has the potential to provide more information regarding why the given school's performance is the same or different from that of the schools in the other school-level latent groups.

For reporting purposes, we define each of the other groups as reference groups. These can be described by considering the characteristics of each group. As mentioned earlier, the school-level latent classes are composed of student-level latent classes. In the MMixIRTM, these differ with respect to the proportion of students-level latent classes within each school-level latent class. In addition, both student-level and school-level latent classes can be described with respect to the demographic variables for both students and schools. In this sense, a reference group at the school-level can be described by considering the proportion of student-level classes and by the particular composition of demographic variables and ability.

Chapter 3

METHODS

The Rasch model, the multilevel IRT model and the MixIRTM are each special cases of the proposed MMixIRTM. If no student-level and school-level mixtures are detected, the MMixIRTM reduces to the Rasch model or multilevel Rasch model. Further, if no school-level latent classes are detected, then the MMixIRTM reduces to a MixIRTM. In this study, the Rasch model, the multilevel IRT model, and the MixIRTM were fitted in the process of model selection. In this chapter, the estimation methods are described for the multilevel IRT model, the MixIRTM, and the MMixIRTM for binary responses. As suggested above, the Rasch model can be considered as a special case of the MixIRTM for a one-group solution, so its estimation is not described here.

3.1 Estimation

3.1.1 A Bayesian Approach to Estimation of a Multilevel Item Response Theory Model

Browne and Draper (2006) list the following programs currently available with specific algorithm for estimation of multilevel models: Maximum likelihood estimation (MLE) Fisher-scoring is implemented in the computer program VARCL (Longford, 1987); the computer program MLwiN (Goldstein, 1986, 1989; Rasbash et al., 2005) implements a number of algorithms including MLE via iterative generalized least squares (IGLS), restricted IGLS for Gaussian data, and marginal quasi-likelihood (MQL) and penalized quasi-likelihood (PQL) algorithms for dichotomous data; empirical-Bayes estimation via the EM algorithm is implemented in the computer program HLM (Raudenbush et al., 2005); and a fully-Bayesian

algorithm is implemented in the computer program MLwin (Goldstein, 1986, 1989; Rasbash et al., 2005).

The PQL estimation method is known to underestimate model parameters for dichotomous data with small sample sizes (Rodriguez & Goldman, 1995; Goldstein & Rashbash, 1996). Rodriguez and Goldman (1995) found that estimates of fixed effects produced by the MQL method had larger bias than estimates produced by standard logistic regression, whether or not the hierarchical nature of the data was accounted for in the model. Goldstein and Rashbash (1996) found PQL estimation to provide biased results albeit less biased than MQL. Browne and Draper (2006) compared random-effects logistic regression (RELR), Bayesian and likelihood-based methods for fitting variance-components. Results with RELR can be considered as similar to those which would be obtained from IRT. Browne and Draper found that PQL for Bernoulli outcomes used for estimating variances of random-effects (e.g., the variance of ability in IRT) with a three-level RELR model performed poorly with respect to bias and convergence. This can be a problem when estimating multilevel IRT models for tests scored dichotomously.

Adams, Wilson, and Wu (1997), Cheong and Raudenbush (2000), and Mislevy (1987) estimated hierarchical models using large-sample approximation or empirical Bayes estimation. These methods introduce constraints that depend on normal distribution theory on the minimum allowable sample size and also on the extent to which sparse data sets affect model estimates in a hierarchical structure (Maier, 2001). Multilevel IRT models also are complex, requiring complex integrations to obtain maximum likelihood solutions. Bayesian approaches do not rely on normal approximations and may permit easier solution strategies that produce unbiased estimates and eliminate computation of complex integrations (Maier, 2001). In addition, Bayesian methods handle missing data relatively well within the parameter estimation scheme (Patz & Junker, 1999a, 1999b; Maier, 2002).

Gibbs Sampling and Markov chain Monte Carlo (MCMC) estimation in general provide flexibility in estimation of multilevel IRT models (Patz & Junker, 1999a, 1999b). Maier

(2001, 2002) described a fully Bayesian approach using MCMC estimation. Fox and Glas (2001) describe a fully Gibbs sampling approach for a multilevel two-parameter normal ogive IRT model and a two-level HLM. In this study, we exploit the fully Bayesian approach for estimation of general MMixIRTM.

Specifying a Model in WinBUGS

Scale anchoring for Multilevel IRT Model. Combining Equations (2.8) - (2.12), the multilevel IRT model can be described as follows:

$$\log\left[\frac{P_i(\theta_{jt})}{1-P_i(\theta_{jt})}\right] = \theta_{jt} - \beta_i = (\gamma_{00} + u_{jt} + \nu_t) - \beta_i, \tag{3.1}$$

where u_{jt} follows $N(0,\tau)$ and ν_t follows $N(0,\zeta)$. Since the γ_{00} is the average mean of ability for the school t, ν_t can be considered, following $N(\gamma_{00},\zeta)$. The metric was anchored based on ability by re-parameterizing u_{jt} into $\sqrt{\tau}\cdot\eta_{jt}$, and setting $\eta_{jt} \sim N(0,1)$.

Priors and Posterior Distributions. The following priors and hyper-priors were used to estimate the parameters of the multilevel IRT model in this study:

$$\begin{aligned}
\eta_{jt} &\sim \text{Normal } (0,1) \\
\nu_t &\sim \text{Normal } (0,\zeta) \\
\beta_i &\sim \text{Normal } (0,1) \\
\zeta &\sim \text{Gamma}(0.1, 0.001) \\
\tau &\sim \text{Normal}(0,1)I(0,) \\
\gamma_{00} &\sim \text{Normal } (0,1)
\end{aligned}$$

where $I(0,)$ indicates that observations of τ were sampled above zero.

The likelihood function for the multilevel IRT model is as follows:

$$L(\theta_{jt}) = \prod_{i=1}^{I}\prod_{j=1}^{J} P(y_{ijt}=1|\theta_{jt})^{u_{ij}} \cdot \{(1-P(y_{ijt}=1|\theta_{jt})\}^{1-u_{ij}}, \tag{3.2}$$

where u_{ij} is dichotomously scored responses as 0 and 1. Note that school identification index ts can be deleted in the likelihood function because student identification index js are nested within the ts.

The joint posterior distribution of

$$S = \{\eta_{jt}, \nu_t, \beta_i, \zeta, \tau, \gamma_{00}\} \tag{3.3}$$

can be written as

$$P(S|U) \quad \propto \quad L(\theta_{jt})P(\eta_{jt}|\tau)P(\tau)P(\nu_t|\gamma_{00}, \zeta)P(\gamma_{00})P(\zeta)P(\beta_i) \ . \tag{3.4}$$

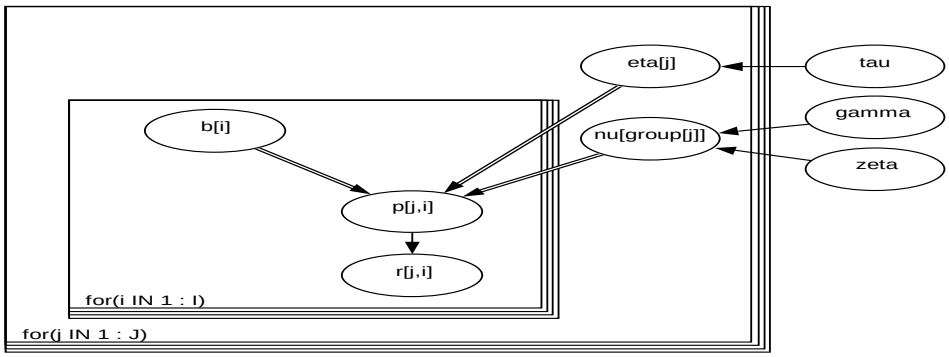
The conditional distribution of each of the individual parameters is derived then from the joint posterior distribution.

Sampling in WinBUGS. WinBUGS includes a graphical modelling facility, called DoodleBUGS, that can be used to graphically represent the full conditional distribution for the multilevel IRT. The resulting graph can also be used in WinBUGS to generate the code to estimate the model. Figure 3.1 shows a graphical model for the multilevel IRT using DoodleBUGS that can be used to generate code which will provide the same result as that in the Appendix A.

The processing in WinBUGS proceeds by sampling all nodes starting at the outer edge of the diagram with the hyperparameters (i.e., *tau*, *gamma*, and *zeta*) and working inwards in the diagram to the $p[j, i]$. $eta[j]$ is the variable name used in the program code for η_j (Note that t was deleted in (3.3)), $b[i]$ is used for b_i, $nu[group[j]]$ is ν_t, *tau* is used for τ, *gamma* is the index for γ_{00}, and *zeta* is the variable name used in the program code for ζ. A solid arrow indicates a stochastic dependence and a hollow arrow indicates a logical function. From the diagram, it can be seen that $eta[j]$ depends on *tau*, and $nu]group[j]]$ depends on *gamma* and *zeta* with $p[j, i]$ being a logical function of b, eta, and $nu]group[j]]$.

Once the model is specified, including all priors and distributions, WinBUGS can determine the necessary sampling methods directly from the graphical structure. The form of the full conditional distributions for γ_{00} and ζ is conjugate, so that direct sampling using standard

Figure 3.1: Graphical Representation of Multilevel IRT



algorithms is done. In addition, the form of the full conditional distributions of $\eta_{jt}, \nu_t, \beta_i, \zeta$, and τ are log-concave distribution, so that derivative-free adaptive rejection sampling is carried out. The truncated version of the normal distribution of τ is also log-concave.

Monitoring Convergence. To determine when convergence was attained, the Gelman and Rubin (1992) statistic, implemented in WinBUGS, was run for two chains using widely disparate initial values for each of the parameters. For example, the initial values for the difficulty parameters, β, were set at -2 and 2, respectively, for the two chains, for the standard deviation of ability, τ, the initial values were 1 and 2, respectively, for the two chains. In addition, the autocorrelation plots were examined. The Geweke (1992), Heidelberger and Welch (1993), and the Raftery and Lewis (1992) methods also were run and analyzed for a single chain.

3.1.2 Bayesian Estimation of Mixture IRT Models

Estimation of MixIRTMs has been accomplished using a number of different algorithms. Estimation using conditional maximum likelihood estimation (CMLE) has been demonstrated for the Rasch family of models (Rost, 1990). The CMLE algorithm is also implemented in the computer program WINMIRA (von Davier, 2001). The MRM and the mixture 2-parameter IRT model also can be fitted with the computer programs Latent GOLD 4.0 (Vermunt & Magidson, 2005) and Mplus (Muthén & Muthén, 2006). MRM formulation in the Latent GOLD 4.0 is based on the conditional response probabilities expressed in terms of log-linear parameters. Latent GOLD uses both the EM and the Newton-Raphson algorithms. In practice, the estimation process starts with a number of EM iterations. When the estimate is judged to be close enough to the final solution, the program switches to the Newton-Raphson algorithm (Vermunt & Magidson, 2005). MRM estimation is also possible in the computer program Mplus (Muthén & Muthén, 2006). Formulation in Mplus is based on conditional response probabilities expressed using a logistic function. Maximum likelihood optimization for the mixture model in Mplus is done in two stages. Muthén and Muthén recommended that optimization first be carried out by using several different randomly specified sets of starting values. The starting values with the highest log-likelihood from among these runs is then used as the starting values in the final stage optimizations. The second (i.e., the final) stage for optimization is carried out using the default optimization settings for TYPE=MIXTURE. When ALGORITHM=INTEGRATION is specified, Mplus employs a maximum likelihood estimator with robust standard errors using a numerical integration algorithm.

Several practical problem arise in MLE for mixture models (Frühwirth-Schnatter, 2006). First, it may be difficult to find the global maximum of the likelihood numerically. Convergence will often fail when the sample size is small or the mixtures are not well-separated (Finch et al., 1989). Second, the likelihood function of mixture models is unbounded and can have many spurious local modes (Kiefer & Wolfowitz, 1956). Third, the provision of standard errors is not straightforward with MLE of finite mixture models, in particular when using

an EM algorithm. Finally, as McLachlan & Peel noted, the sample size has to be very large to apply the asymptotic theory of maximum likelihood for mixture models.

A Markov chain Monte Carlo (MCMC) estimation algorithm, however, has been found to be useful for estimating mixture distributions (Diebolt & Robert, 1994; Robert, 1996), including MixIRTMs (Bolt, Cohen, & Wollack, 2001, 2002; Cho, Cohen, & Kim, 2006; Cohen & Bolt, 2005; Cohen, Cho, & Kim, 2005; Samuelsen, 2005; Wollack, Cohen, & Wells, 2003). A Bayesian approach can be useful for estimating finite mixture models, in fact, because use of a proper prior will usually operate to smooth the likelihood function of the mixture model (Frühwirth-Schnatter, 2006). This, in turn, serves to reduce the possibility that a spurious model might be selected in cases where an approach such as an EM algorithm would possibly lead to a degenerate solution. Further, parameter uncertainty is easier to address, since the complete posterior is available (Frühwirth-Schnatter, 2006). Also, since a Bayesian estimation algorithm does not rely on asymptotic normality, it will yield valid inferences in situations in which regularity conditions are violated, such as small data sets and mixtures with small mixture proportions (Frühwirth-Schnatter, 2006).

Three issues are addressed below for mixture modeling: Label switching, model selection, and use of priors.

Label switching

A potentially serious problem in Bayesian estimation of mixture models is that of identification. Identification of a model implies that different parameter estimates yield different likelihood values. In a Bayesian context, non-identification would mean that the posterior is the same for different values of the parameter.

Identifiability for a mixture distribution is defined as follows (McLachlan & Peel, 2000). Let

$$f(y_j; \Psi) = \sum_{g=1}^{G} \pi_g \cdot f_g(y_j; \theta_g) \tag{3.5}$$

and

$$f(y_j; \Psi^*) = \sum_{g=1}^{G^*} \pi_g^* \cdot f_g(y_j; \theta_g^*) \tag{3.6}$$

be any two members of a parametric family of mixture densities. y_j is the observed value of the random vector Y_j, Ψ is a parameter in the parameter space, Ω, θ_g and θ_g^* are the vectors of unknown parameters, and π_g and π_g^* are mixing proportions. This class of finite mixtures is said to be identifiable for $\Psi \in \Omega$ if

$$f(y_j; \Psi) \equiv f(y_j; \Psi^*) \tag{3.7}$$

if and only if $G = G^*$.

Bayesian analysis can sometimes help with model identification by the use of priors. For example, priors imposing just enough information can be added to uniquely determine the parameter values (Vermunt & Magidson, 2005).

Cho et al. (2006) describe two types of label switching in the MixIRTM. The first type is that which is commonly referred to as label switching in mixture models: It occurs across iterations within a single MCMC chain. The second type occurs when the latent classes switch among replications (as in a simulation study) or for different initial values. This latter kind occurs in both maximum likelihood and Bayesian solutions.

The first type of label switching can be a serious problem in Bayesian estimation because the labels of the mixtures can change within the MCMC chain on different iterations. That is, the meaning of the latent classes can simply switch. This type of label switching occurs essentially because there is a lack of sufficient information that can be used by the algorithm to discriminate between the latent groups of mixture models belonging to the same parametric family (McLachlan & Peel, 2000). The root of this type of label switching is that the likelihood is the same for all permutations of parameters (Stephens, 2000). It can be seen, when distinct jumps occur in the traces of a parameter over the course of the MCMC chain, and also when the plot of the density for the parameter has multiple modes (Stephens, 2000). If multiple modes do exist for any of the distributions of parameters, then label switching

is present with the result that interpretation of the posterior of the parameter is distorted. This type of label switching also can be detected by examining the marginalized posterior distribution of each parameter.

The second type of label switching sometimes arises in a MCMC chain. This is not the usual within-chain label switching but rather occurs among different chains. The result of this form of label switching can cause confusion in interpretation of results because the latent classes have a different order in each chain. This kind of label switching can easily be observed in a simulation study because the generating parameters are available and can be compared with the parameter estimates to determine which labels should be applied to each of the latent classes. This second form of switching of latent class labels among replications is not uncommon and was observed in the simulation study described in Cho et al. (2006). Evidence for this second type of label switching was found by comparing results for multiple chains with different starting values. Group memberships were compared for each chain with the generating parameters to determine whether class labels had switched among the different replications. Checking convergence when this kind of label switching occurs can be a problem with more than two chains for empirical data since the true group memberships are not known. The approach used here was to try to match the estimated group memberships by comparing estimated group ability means and proportions across the different chains.

The following suggestions may be useful in helping to avoid the first type of the label switching for the empirical data described above. One way to limit label switching is to impose an identifiability constraint on the parameter space that can be satisfied by only one of the possible permutations for a given parameter. Although there are many choices that could be made for identifiability constraints for any given data set, not all will be effective in removing the symmetry in the posterior distribution (Stephens, 2000). Furthermore, setting constraints is somewhat problematic in that such constraints may possibly distort the final estimate (Congdon, 2003). Stephens (2000) also showed, however, that artificial identifiability constraints generally do not solve this type of label switching. Instead, Stephens suggested

a relabelling algorithm, based on a decision theoretic approach, in which first a loss function is defined, and then the estimates are chosen that minimize the posterior expected loss for that function.

Model Selection

When the number of components in a model is unknown, the parameter space is ill-defined and of infinite dimension (McLachlan & Peel, 2000). One approach in Bayesian estimation of mixture models is to consider the number of latent mixtures, g, to be an unknown parameter with a prior distribution. Richardson and Green (1997) describe an approach in which the number of latent mixtures is an unknown parameter that is to be estimated. The usual approach in mixture modeling, and the one assumed in this paper, however, is to fit a range of mixture models each with a different number of latent classes, and then to consider the selection of a model according to some theoretical rationale often including use of some appropriate statistical criteria.

Li et al. (2006) examined several model selection indices for use in Bayesian estimation of dichotomous MixIRTMs: Akaike's information coefficient (AIC), Bayesian information coefficient (BIC), deviance information coefficient (DIC), pseudo-Bayes factor (PsBF), and posterior predictive checks (PPP). The five indices provided somewhat different recommendations for a set of real data. Results from a simulation study indicated that BIC selected the generating model well under all conditions simulated and for all three of the dichotomous MixIRTMs (i.e., the MRM, the mixture 2-parameter logistic model, and the mixture 3-parameter logistic model) considered. PsBF was almost as effective. AIC and DIC tended to select the more complex model, and PPP was essentially ineffective for this use.

AIC and BIC were used in this study for selection of the model with the correct number of mixtures. Since the ζ_{jg}^{t} may be different at each iteration in sampling, it is necessary to monitor the likelihood at each iteration. The definitions of AIC and BIC in this study are

as follows:

$$AIC = -2D^t + 2m, \tag{3.8}$$

$$BIC = -2D^t + m \log n \tag{3.9}$$

where m is the number of parameters, and n is the number of observations.

The model selection strategy taken in this study is one in which the candidate models, each describing different numbers of mixtures, are run in parallel. Information is then accumulated over iterations to provide a probability that a specific model is selected by AIC and BIC (Congdon, 2003). This approach compares the averages of the AIC and BIC fit measures over iterations in a MCMC run following burn-in.

The Use of Priors

Problems which may arise in estimation of MixIRTMs can sometimes be avoided or at least ameliorated to some extent by use of a Bayesian estimation algorithm. Information in the form of priors can be used to direct the algorithm away from extremes which might otherwise be approached during estimation. Bayesian algorithms make use of priors to constrain or guide the estimation of model parameters. The effect of the prior is greater, given the data, to the extent it imposes a stronger constraint on the algorithm. When the information used by the estimating algorithm comes solely from the data, or when priors are uninformative, then a Bayesian estimate would be comparable to that obtained from a non-Bayesian algorithm such as a MLE algorithm.

A fully Bayesian estimation of the MRM requires that priors are set for ability, item parameters, and mixture probabilities. Cho et al. (2006) followed standard practice by setting the prior distribution for ability as normal with mean 0 and standard deviation 1 for the reference group. Normal prior distributions on the item parameters were also used to improve the stability of the fitting process. These priors have been reported to yield reasonable estimates of the item parameters (Johnson & Albert, 1998).

The form of the prior is an important issue in Bayesian estimation. Improper priors will yield improper posterior distributions. If an improper prior is adopted for the mixture component parameters, then the posterior distribution will be improper. One approach to avoiding improper posterior distributions in mixture models is to use conjugate priors (Diebolt & Robert, 1994). Cho et al. (2006) implemented an MCMC method for the MRM and compared a Dirichlet prior, a Dirichlet process with stick-breaking prior, and a multinomial regression model on the probability of mixture for model parameter recovery. There were no noticeable differences between the Dirichlet prior and the Dirichlet process with stick-breaking prior. For the two-group simulation, recovery was nearly the same for both priors and the logistic regression model with a covariate. For the three-group simulation, however, recovery of item difficulties and group memberships improved with the use of the logistic regression model with a covariate. The four-group condition was simulated as having one group that was a guessing group. Recovery for this condition was less consistent than for the three-group condition. The Dirichlet prior was used for MRM in this study.

Specifying a Model in WinBUGS

Scale Anchoring and Linking. Scales were linked so parameter estimates from different latent classes were placed on the same metric. This permits direct comparisons of the parameter estimates among the different latent classes. The usual way of identifying the metric for the MRM is to impose the restriction $\sum_i b_{ig} = 0$ on the difficulty parameters within each class (Rost, 1990). Using this approach is effective for identification and also for scale anchoring in the MRM. For other models, an additional step is needed to link the metrics of the different latent classes.

In this study, the metric was anchored by selecting one of the latent classes in the MRM as a reference group, and setting the ability distribution as $logit[P(y_{ijg} = 1|g, \eta_{jg})] = \theta_{jg} - \beta_{ig} = \sigma_g \cdot \eta_{jg} - \beta_{ig}$ with the following priors:

$$\eta_{jg} \sim \text{Normal}(\mu_g, 1), \quad g = 1, \ldots, G$$

$$\begin{aligned} \mu_1 &= 0, \mu_2 \sim \text{Normal}(0,1) \\ \sigma_g &\sim \text{Normal}(0,1)I(0,). \end{aligned}$$

where $\eta_{jg} = \frac{\theta_{jg}}{\sqrt{\sigma_g^2}}$. For anchoring, the mean of η was set to 0 by selecting one of the latent classes in the model as the reference group. In this study, the metric of the first latent class was selected as the base or reference metric. The means for the remaining latent classes were set to μ_g, that is, they were estimated relative to the metric of the reference group. Thus, the mean and variance of the ability distributions of the other groups were estimated relative to the $N(0,1)$ scale of the reference group.

Priors and Posterior Distributions. The following priors and its hyper-priors were used to estimate the parameters of the MRM:

$$\begin{aligned} g &\sim \text{Multinomial } (1, \pi_g[1:G]) \\ \eta_j | G = g &\sim \text{Normal } (\mu_g, 1), \quad j = 1, \ldots, N \\ b_{ig} &\sim \text{Normal } (0,1), \quad i = 1, \ldots, n, \quad g = 1, \ldots, G \\ \sigma_g &\sim \text{Normal}(0,1)I(0,), \quad g = 1, \ldots, G \\ \mu_g &\sim \text{Normal } (0,1), \quad g = 2, \ldots, G, \mu_1 = 1, \end{aligned}$$

where $I(0,)$ indicates that observations of a were sampled above zero.

Two things were used on π_g: a Dirichlet prior as a conjugate prior and a multinomial logistic regression model with a covariate for π_g. An mildly informative prior was used on item difficulty. This prior was set at $b_{ig} \sim N(0,1)$ across items and classes. The use of diffuse priors failed to provide enough bound on the item difficulty and standard deviation of ability parameters for MRM. As a result, mildly informative prior was used. The use of mildly informative priors was done to provide rough bounds on the parameters of the model and to make fitting procedures more stable (Bolt, Cohen, & Wollack, 2001, 2002; Cohen & Bolt, 2005; Cohen, Cho, & Kim, 2005; Samuelsen, 2005; Wollack, Cohen, & Wells, 2003).

The likelihood function for the MRM is as follows:

$$L(g, \theta_{jg}) = \quad \prod_{i=1}^{I} \prod_{j=1}^{J} [\{\sum_{g=1}^{G} \pi_g \cdot P(y_{ijg} = 1 | g, \theta_{jg})\}^{u_{ij}}$$

$$\cdot \{\sum_{g=1}^{G} \pi_g \cdot (1 - P(y_{ijg} = 1|g, \theta_{jg})\}^{1-u_{ij}}]^{\zeta_{jg}^t} \ , \tag{3.10}$$

where u_{ij} is dichotomously scored as 0 =incorrect and 1 =correct, $\zeta_{jg} = 1$ if examinee j is from mixture g and 0 otherwise.

With $\theta_{jg} = \sqrt{\sigma_g^2} \cdot \eta_{jg} = \sigma_g \cdot \eta_{jg}$, the joint posterior distribution of

$$S = \{\eta_j, \mu_g, \sigma_g, b_{ig}, g, \pi_g\} \tag{3.11}$$

the Dirichlet prior as a conjugate prior can be written as

$$P(S|U) \quad \propto \quad L(g, \theta_j)P(\eta_j|\mu_g)P(\mu_g)P(\sigma_g)P(b_{ig})P(g|\pi_g)P(\pi_g). \tag{3.12}$$

For a multinomial logistic regression model, the joint posterior distribution of

$$S = \{\eta_j, \mu_g, \sigma_g, b_{ig}, g, \pi_g, \gamma_{0g}, \gamma_{pg}\} \tag{3.13}$$

can be written as

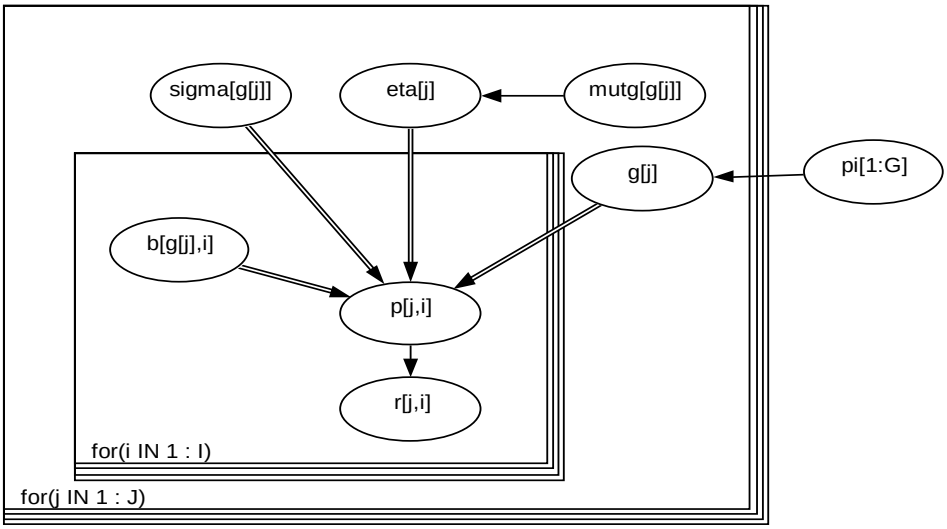
$$P(S|U) \quad \propto \quad L(g, \theta_j)P(\eta_j|\mu_g)P(\mu_g)P(\sigma_g)P(b_{ig})P(g|\pi_g)P(\gamma_{0g})P(\gamma_{pg}). \tag{3.14}$$

Since the Rasch model can be considered as a special case of a MRM with a one-group solution, the above procedure can be used. The WinBUGS code for this model is given in Appendix B.

Sampling in WinBUGS. As noted above, WinBUGS includes a graphical modeling facility called DoodleBUGS that can be used to graphically represent the full conditional distributions for the MRM. DoodleBUGS can also render WinBUGS code based on the diagram to estimate the model based. Figure 3.2 shows a graphical model for the MRM using DoodleBUGS that can be used to generate code which will provide the same results as the code in Appendix C. For illustration purposes, a Dirichlet prior on the probabilities of the mixtures is shown in Figure 3.2.

The order of the processing in WinBUGS can be found in the same DoodleBUGS diagram. Processing (as described in Figure 3.2) proceeds by sampling all nodes starting at the outer

Figure 3.2: Graphical Representation of MRM



edge of the diagram with the hyperparameters and working inwards in the diagram to the $p[j, i]$. $eta[j]$ is the variable name used in the program code for η_j, $mutg[g[j]]$ is used for μ_g, $sigma[g[j]]$ is σ_g, $b[g[j], i]$ is used for class-specific item difficulty b_{ig}, $g[j]$ is the index for group membership for person j, and $pi[g]$ is the variable name used in the program code for π_g. A solid arrow indicates a stochastic dependence and a hollow arrow indicates a logical function. From the diagram, it can be seen that $eta[j]$ depends on $mutg[g[j]]$, and

$g[j]$ depends on $pi[g]$ with $p[j,i]$ being a logical function of $sigma[g[j]]$, $eta[j]$, $b[g[j],i]$, and $g[j]$.

Once the model is specified, including all priors and distributions, WinBUGS can determine the necessary sampling methods directly from the graphical structure. The form of the full conditional distributions for μ_g and π_g are both conjugate (i.e., normal and Dirichlet distribution, respectively), so that direct sampling using standard algorithms is done. The form of the full conditional distribution of g is a discrete distribution with a finite upper bound so that the inversion method is implemented. In addition, the forms of the full conditional distributions of η_j, σ_g, and b_{ig} are log-concave distribution, so that derivative-free adaptive rejection sampling is carried out. The truncated version of the normal distribution of σ_g is also log-concave.

A common MCMC strategy is to sample a class membership parameter for each observation at each stage of the Markov chain (Robert, 1996). In this study, the algorithm implemented the following strategy: A class membership parameter, $g = 1, 2, \cdots, G$, was sampled for each examinee j along with a continuous latent ability parameter η_j. The group membership parameter was determined as the mode of gs across iterations following burn-in.

Monitoring Convergence. In Cho et al. (2006), some of the chains had not converged by the time they had reached 100,000 iterations. This occurred for all the 10-item conditions for each sample size and for the 30-item and 50-item conditions with 360 examinees in that study. These results indicate that the data were not sufficiently informative to obtain stable parameter estimates for these conditions. The results further suggest that test lengths for the MRMs simulated should be at least 30 items and sample sizes should be around 1,000 examinees. For those conditions in Cho et al. that did converge, results suggested a conservative burn-in of 7,000 iterations would be appropriate for all conditions for which convergence was obtained. Based on the results from these procedures, Cho et al. adopted a common conservative burn-in of 7,000 burn-in iterations and 8,000 post burn-in iterations for all remaining conditions.

Mixture models in general are likely to have multiple local maxima of the likelihood (Muthén et al., 2002). Maximizing a likelihood yields a solution that provides a local maximum only within a restricted set of parameter values rather than globally over all possible combinations of parameter values. The usual method used to check whether the model is identified or whether there is merely a local solution is to run the model with different starting values (McLachlan & Peel, 2000). Observing the same log likelihood obtained from multiple set of initial values, for example, increases confidence that the solution is not a local maximum. This is similar to the checking of convergence of two or more chains by using markedly different starting values in a Bayesian solution. Once convergence has been reached, a Bayesian solution may prevent the possibility both of non-identification and local maxima when estimating mixture models.

3.1.3 Bayesian Estimation of the Multilevel Mixture IRT Model

The MMixIRTM parameters were estimated using a Bayesian algorithm. A MCMC Gibbs sampling algorithm was written using WinBUGS 1.4 (Spiegelhalter, et al., 2003) to estimate the MMixIRTM with covariates model described above (The WinBUGS program code used is given in Appendix D and E.) Considered below are issues to be addressed in the MCMC solution.

Label Switching

There are three possible types of label switching that can arise in a MMixIRTM. Label switching within a chain is the first type, and label switching across chains is the second type. The third type of label switching is a variant of Type II switching in which student-level latent classes switch within a school-level latent class. Each type of label switching is described below.

The lack of identifiability causes problems in a Bayesian framework where the posterior distribution is used to make inferences regarding model parameters. This problem is harder

to deal with for a MMixIRTM because of the multilevel structure of mixtures. Employing the notion of the hierarchical mixtures-of experts (HME) model suggested in Jordan and Jacobs (1992), the MMixIRTM can be viewed as having a two-layer mixture structure. The higher-level is the school level and mixing proportions in the latent classes are π_k. At the lower-level, the student-level, latent class have mixing proportions $\pi_{g|k}$. Vermunt (in press a) described the model identification issue for the multilevel mixture model as follows: A necessary condition for identification is that the higher-level model has the structure of an identifiable latent class model. Separate identifiability of the lower part of the model is a sufficient condition but not always necessary when the number of higher-level classes is larger than 1. A necessary condition for identification is that the π_k for the K latent classes should be identified. As noted earlier, label switching can be seen when distinct jumps occur in the traces of a parameter and when the density for the parameter has multiple modes (Stephens, 2000). If multiple modes do not exist for the π_k, it can be concluded that there is no the first type label switching and a necessary condition for identification is satisfied.

The second type of label switching arises in a MCMC chain as described MixIRTM. This kind of label switching may happen within each school-level mixture in the MMixIRTM since the student-level proportion is modelled within a school-level mixture. That is, the labelling of student-level membership is different for each school-level mixture, if this second type of label switching arises within a school-level mixture. This type of label switching is called the third type of label switching in this study. In a simulation study, the parameter estimates can be compared with the generating parameters to determine which labels should be applied to each of the latent classes for each school-level mixture. In an empirical data sets, group memberships can be matched across chains and across school-level mixtures by looking at the patterns of the means of ability, mixture proportions, and difficulty.

Model Selection

As described (above) in the section on model selection for Bayesian estimation of MixIRTM, AIC and BIC were used for the model selection indices.

Specifying a Model in WinBUGS

Scale Anchoring and Linking for the MMixIRTM. Linking of scales is necessary in order to make comparisons among the different latent groups in the model. In this study, we anchor the metric with respect to ability distribution. This procedure is done by re-parameterizing θ_{jtgk} into $\sqrt{\sigma^2_{gk}} \cdot \eta_{jtgk}$. That is, θ_{jtgk} and η_{jtgk} are set, respectively, as follows:

$$\begin{aligned} \theta_{jtgk} &\sim \text{Normal}(\mu_{gk}, \sigma^2_{gk}), \\ \eta_{jtgk} &\sim \text{Normal}(\mu_{gk}, 1), \end{aligned}$$

where $\eta_{jtgk} = \frac{\theta_{jtgk}}{\sqrt{\sigma^2_{gk}}}$. For metric anchoring, the mean of η_{gk} is set to 0 for the reference group (in this study, $\mu_{11} = 0$). The means for the remaining latent classes are set to μ_{gk}, that is, they are to be estimated. Thus, the mean and variance of the distributions of the other groups are estimated relative to the $N(0, 1)$ scale of the reference group in terms of η_{jtgk}. That is, the procedure can be described as follows:

$$logit[P(y_{ijtgk} = 1|g, k, \eta_{jtgk})] = \sqrt{\sigma^2_{gk}} \cdot \eta_{jtgk} - b_{igk}. \tag{3.15}$$

DIF analysis with MMixIRTM is the situation where the same set of items are administrated across unknown groups, i.e., latent classes, g and k. That is, each (latent) group of examinees responded to the same set of items. Given the particular items used, one can think of every item on the scale as being a potential anchor item to be used in estimating an appropriate link. This is similar to a common-item internal anchor nonequivalent groups linking design. However, in the MMixIRTM, class-specific item difficulties as well as group memberships g and k are estimated simultaneously. For comparisons of the item difficulties

across latent classes, g and k, The b_{Tigk} are calculated for each classes g and k with the $\sum_i^I b_{igk} = 0$.

Priors and Posterior Distributions. The following priors were used for the binary response version of the MMixIRTM:

$$\begin{aligned}
g &\sim \text{Categorical } (\pi_{g|k}[1:G]) \\
k &\sim \text{Categorical } (\pi_k[1:K]) \\
\eta_{jt}|G=g,K=k &\sim \text{Normal } (\mu_{gk},1), \quad j=1,\ldots,J, t=1,\ldots,T \\
\mu_{gk} &\sim \text{Normal } (0,1), \quad g=2,\ldots,G, k=2,\ldots,K, \mu_{11}=1 \\
\sigma_{gk} &\sim \text{Normal}(0,1)I(0,), \quad g=1,\ldots,G, \\
\beta_{igk} &\sim \text{Normal } (0,1), \quad i=1,\ldots,I, g=1,\ldots,G,
\end{aligned}$$

where $I(0,)$ indicates that observations of σ were sampled above 0. An mildly informative prior on item difficulty was set at $b_{igk} \sim N(0,1)$ for items across classes. The use of diffuse priors failed to provide enough bound on the item difficulty and standard deviation of ability parameters for MMixIRTM. As a result, mildly informative prior was used. The use of such priors was done to provide rough bounds on the parameters of the model and to make fitting procedures more stable (Bolt, Cohen, & Wollack, 2001, 2002; Cohen & Bolt, 2005; Cohen, Cho, & Kim, 2005; Samuelsen, 2005; Wollack, Cohen, & Wells, 2003).

A multinomial logistic regression with a covariate model was used for representing the student-level mixtures conditional on the school-level mixture (i.e., $\pi_{g|k}$). The following models were used:

$$\pi_{g|k,W_j} = \frac{\exp(\gamma_{0gk} + \sum_{p=1}^P \gamma_{pg} W_{jp})}{\sum_{g=1}^G \exp(\gamma_{0gk} + \sum_{p=1}^P \gamma_{pg} W_{jp})}, \tag{3.16}$$

and

$$\pi_{g|k,W_j} = \frac{\exp(\gamma_{0gk} + \sum_{p=1}^P \gamma_{pgk} W_{jp})}{\sum_{g=1}^G \exp(\gamma_{0gk} + \sum_{p=1}^P \gamma_{pgk} W_{jp})}. \tag{3.17}$$

The difference between the first model and the second model is that the first model contains a covariate effect (γ_{pg}) for school-level mixtures, whereas the second model contains the

covariate effect (γ_{pgk}) for school-level mixtures. Priors for $\gamma_{0gk}, \gamma_{pg}$, and γ_{pgk}, were set to $N(0,1)$.

For the prior of $\pi_{g|k}$, a Dirichlet distribution can be used as the conjugate prior of the parameters of the multinomial distribution:

$$\frac{\Gamma(\sum_g \alpha_g)}{\prod_g \Gamma(\alpha_g)} \cdot \prod_g^G \pi_{g|k}^{\alpha_g - 1}, \tag{3.18}$$

where $\sum_{g=1}^{G} \pi_{g|k} = 1$ for all school-level latent class ks with the proportion of π_k, and G indicates the number of student-level latent classes.

One way to sample $\pi_{g|k}$ from the Dirichlet distribution is to sample G independent random variables $\pi^*_{g|k}$ from the Gamma distribution, $Gamma(\alpha_g, 1)$, $g = 1, \cdots, G$ normalizing

$$\pi_{g|k} = \frac{\pi^*_{g|k}}{\sum_{g=1}^{G} \pi^*_{g|k}}, \tag{3.19}$$

for each k.

A second way to sample $\pi_{g|k}$ from the Dirichlet distribution is to set α_g directly. The parameters α_g can be interpreted as "prior observation counts" for events governed by $\pi_{g|k}$ (Gelman et al., 2003). The density is proper if all $\alpha_{g|k}$ are positive. The density of the Dirichlet distribution is improper if all $\alpha_{g|k}$ are 0. When $\alpha_{g|k} \to 0$, the distribution becomes noninformative. The order of mixtures in a mixture density is arbitrary. Consequently, the usual procedure is to select priors that reflect this information, such that they are invariant to relabelling of the mixtures (Frühwirth-Schnatter, 2006). Thus, the hyperparameters of the prior in this study were assumed to be the same, yielding invariant priors. That is, the means of all the $\pi_{g|k}$ priors were the same in this study. The marginal distribution of $\pi_{g|k}$ is a Beta distribution, therefore, the mean and variance of the $\pi_{g|k}$ are respectively as follows: (Frühwirth-Schnatter, 2006):

$$E(\pi_{g|k}) = \frac{\alpha_g}{\sum_g \alpha_g} \tag{3.20}$$

and

$$Var(\pi_{g|k}) = \frac{\alpha_g \cdot (\sum_g \alpha_g - \alpha_g)}{(\sum_g \alpha_g)^2 \cdot (\sum_g \alpha_g + 1)} \ . \tag{3.21}$$

The variances decrease as the parameter values of α_g increase. In this study, $\alpha_g = 1$ as a uniform Dirichlet prior and as $\alpha_g = 4$ were both studied as possible priors. A multinomial logistic regression with a school-level covariate, a Dirichlet prior as a conjugate prior, a Dirichlet prior with the Gamma distribution, and a Dirichlet process with stick-breaking prior also can be used for the school-level probabilities of mixtures (i.e., π_k).

The probability of a school belonging to latent class k, π_k, can be written as

$$\pi_{k|W_t} = \frac{\exp(\gamma_{0k} + \sum_{p=1}^{P} \gamma_{pk} W_{tp})}{\sum_{k=1}^{K} \exp(\gamma_{0k} + \sum_{p=1}^{P} \gamma_{pk} W_{tp})}. \tag{3.22}$$

Priors of γ_{0k} and γ_{pk}, were set to $N(0, 1)$.

The Dirichlet distribution is the conjugate prior of the parameters of the multinomial distribution. It is defined as follows in the computer program WinBUGS 1.4 (Spigelhalter et al., 2003):

$$\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \cdot \prod_k^K \pi_k^{\alpha_k - 1}, \tag{3.23}$$

where $0 < p_k < 1$, $\sum_k p_k = 1$, and K indicates the number of school-level mixtures. In this study, $\alpha_k = 1$ as a uniform Dirichlet prior and $\alpha_k = 4$ were tried for prior investigation.

Another way to sample π_k from the Dirichlet distribution is to sample K independent random variables π_k^* from the Gamma distribution, $Gamma(\alpha_k, 1)$, $k = 1, \cdots, K$ normalizing

$$\pi_k = \frac{\pi_k^*}{\sum_{k=1}^{K} \pi_k^*}, \tag{3.24}$$

for each k.

The Dirichlet process prior (DPP) offers an approach which avoids parametric assumptions and is less impeded by uncertainty about the appropriate number of classes. The DPP method deals with possible clustering in the data by only requiring specification of a maximum conceivable number of latent classes (Congdon, 2003). π_k at any iteration may be constructed by defining a sequence $r_1, r_2, \cdots, r_{K-1}$ of $Beta(1, \alpha)$ random variables, with $r_K = 1$ (Congdon, 2003). The distribution of $Beta(a, b)$ is defined as follows in the computer program WinBUGS 1.4 (Spigelhalter et al., 2003):

$$p_{a-1}(1-p)_{b-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma b}, \tag{3.25}$$

where $0 < p < 1$.

The stick-breaking prior proceeds as follows: Sample

$$r_g \sim Beta(1, \alpha), g = 1, \cdots, G, \tag{3.26}$$

and set

$$\pi_1 = r_1, \tag{3.27}$$

$$\pi_2 = r_2 \cdot (1 - r_1), \tag{3.28}$$

$$\pi_3 = r_3 \cdot (1 - r_2) \cdot (1 - r_1), \tag{3.29}$$

and so on. This is known as a stick-breaking prior since at each stage what is left of a stick of unit length is broken and the length of the broken portion assigned to the current value π_k. At any iteration, some of the K potential classes may be empty. It is possible to monitor the actual number K of non-empty, though differing, classes.

The likelihood function for the binary response version of the MMixIRTM is as follows:

$$\begin{aligned} L(g, k, \theta_{jg}) = \; & \prod_{i=1}^{I} \prod_{j=1}^{J} [\{\sum_{k=1}^{K} \sum_{g=1}^{G} \pi_k \cdot \pi_{g|k} P(y_{ijtgk} = 1 | g, k, \theta_{jtgk})\}^{u_{ij}} \\ & \cdot \{\sum_{k=1}^{K} \sum_{g=1}^{G} \pi_k \cdot \pi_{g|k} P(y_{ijtgk} = 1 | g, k, \theta_{jtgk})\}^{1-u_{ij}}]^{\zeta_{jgk}^t} , \end{aligned} \tag{3.30}$$

where u_{ij} is dichotomously scored responses as 0 and 1, $\zeta_{jgk} = 1$ if the examinee j is from mixtures g and k and $\zeta_{jgk} = 0$ otherwise.

The joint posterior distribution for the use of priors on $\pi_{g|k}$ and π_k,

$$S = \{g, k, \eta_{jt}, \mu_{gk}, \sigma_{gk}, \beta_{igk}, \pi_k, \pi_{g|k}\}$$

can be written as

$$\begin{aligned} P(S|U) \propto \; & L(g, k, \eta_{jt}) P(\eta_{jt} | \mu_{gk}) P(\mu_{gk}) P(g | \pi_{g|k}) \\ & \cdot P(\pi_{g|k}) P(k | \pi_k) P(\pi_k) P(a_{gk}) P(\beta_{igk}) . \end{aligned} \tag{3.31}$$

The joint posterior distribution for the use of a multinomial logistic regression model on $\pi_{g|k}$ and π_k,

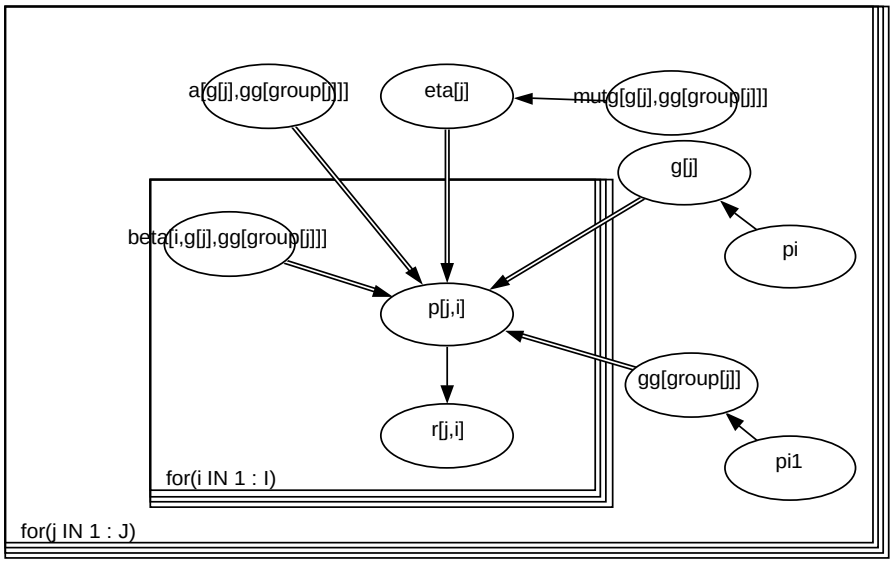
$$S = \{g, k, \eta_{jt}, \mu_{gk}, \sigma_{gk}, \beta_{igk}, \pi_k, \pi_{g|k}, \gamma_{0gk}, \gamma_{pg}, \gamma_{0k}, \gamma_{pk}\}$$

can be written as

$$P(S|U) \propto \quad L(g,k,\eta_{jt})P(\eta_{jt}|\mu_{gk})P(\mu_{gk})P(g|\gamma_{0gk},\gamma_{pg})P(\gamma_{0gk})P(\gamma_{pg})$$
$$\cdot\ P(k|\gamma_{0k},\gamma_{pk})P(\gamma_{0k})P(\gamma_{pk})P(a_{gk})P(\beta_{igk})\ . \tag{3.32}$$

Sampling in WinBUGS. MCMC sampling in WinBUGS is shown below for the use of priors on the probability of mixtures. WinBUGS uses a graphical model to calculate full conditional distribution. Figure 3.3 is shown to represent the graphical model of MMixIRTM with DoodleBUGS, which leads the same code in the Appendix D we used for the estimation.

Figure 3.3: Graphical Representation of MMixIRTM with Priors



The processing in WinBUGS proceeds by sampling all nodes starting at the outer edge of the diagram with the hyperparameters and working inwards in the diagram to the $p[j,i]$. As an example, pi is the variable name used in the program code for $\pi_{g|k}$, $pi1$ is the variable name used in the program code for π_k, $g[j]$ is the index for group membership for student j, $gg[group[j]]$ is the index for group membership for school k, $eta[j]$ is η_j, $a[g[j], gg[group[j]]$ is σ_{gk}, $beta[i, g[j], gg[group[j]]]$ is β_{igk} and $mutg[g[j], gg[group[j]]]$ is μ_{gk}. A solid arrow indicates a stochastic dependence and a hollow arrow indicates a logical function. From the diagram, it can be seen that $eta[j]$ depends on $mutg[g[j], gg[group[j]]]$, $g[j]$ depends on pi, and $gg[group[j]]$ depends on $pi1$. $p[j,i]$ (which is the code in the program for $p(P(y_{ijt} = 1|g, k, \theta_{jgk}))$) is a logical function of $a[g[j], gg[group[j]], beta[i, g[j], gg[group[j]]]$, $eta[j], mutg[g[j], gg[group[j]]], g[j]$, and $gg[group[j]]$.

Once the model is fully specified using the distributions given above, WinBUGS then determines the necessary sampling methods directly from the structure in the diagram. The form of the full conditional distribution of $\mu_{gk}, \pi_{g|k}$, and π_k is a conjugate distribution of the parameters (i.e., normal and Dirichlet distributions), so that in this study, direct sampling was conducted using standard algorithms. The form of the full conditional distribution of g and k is a discrete distribution with a finite upper bound so that the inversion method is implemented. The form of the full conditional distribution of η_j, a_{gk}, and β_{igk} is log-concave, so that WinBUGS uses derivative-free adaptive rejection sampling. The truncated version of the normal distribution of σ_{gk} is log-concave as well.

Starting values are needed for each parameter being sampled in order to define the first state of the Markov chain. The starting values for the remaining model parameters were randomly generated in the WinBUGS software except that the school-level group membership to monitor the trace of it. It was randomly set with the random permutation for both simulation and empirical study.

Monitoring Convergence. The Gelman and Rubin (1992) statistic was run for two chains using widely disparate initial values for each of the parameters (e.g., the initial values

for the difficulty parameters, β_{gk}, were set at -2 and 2, respectively, for the two chains, for the standard deviation of ability, σ_{gk}, the initial values were 1 and 2, respectively, for the two chains). The autocorrelation plot was examined. The Geweke (1992), Heidelberger and Welch (1993), and the Raftery and Lewis (1992) methods also were run and analyzed for a single chain.

3.2 DIF Detection Procedure

In this section, the DIF detection procedure of the general MMixIRTM is described at both student- and school-levels.

Student-level DIF in the General MMixIRTM. As described earlier, DIF at the student-level is defined based on differences in item difficulties within a school-level latent class. In other words, we obtain information about student DIF for each school-level latent class. Since we are concerned with school-level comparisons in this study, we do not define one of the student-level latent classes as a reference group. Further, the student-level DIF measures are based on differences in item difficulties of all pairwise comparisons for binary and for polytomous items as follows:

$$DIFM = \hat{\beta}_{Tigk} - \hat{\beta}_{Tig'k}, \tag{3.33}$$

and

$$DIFM = \hat{\delta}_{Tigk} - \hat{\delta}_{Tig'k}, \tag{3.34}$$

where $\hat{\beta}_{Tigk}$ and $\hat{\delta}_{Tigk}$ are transformed (T) estimated item difficulties for the given k.

For studied item i of the given school-level in latent class k, the following comparisons are of interest:

$$\mathrm{H}_0 : \beta_{Tigk} - \beta_{Tig'k} = 0, \tag{3.35}$$

and

$$\mathrm{H}_0 : \delta_{Tigk} - \delta_{Tig'k} = 0 \ , \tag{3.36}$$

indicating that there is no hypothesized difference in item difficulties between two student-level groups.

School-level DIF in the General MMixIRTM. DIF at the school-level can be determined by comparing the differences between the item difficulties among school-level latent classes. The current procedure used by the College Board is to provide a DIF analysis for each school relative to a single comparison group. In this study, this same approach is used except that there are $K-1$ latent classes, each of which is used as a comparison group.

Each school is classified by the MMixIRTM into to one of the K school-level latent classes. For purposes of reporting to each school, we identify the latent class in which that school is a member and define that latent class as the focal group. For ease of discussion, we refer to the given school of interest as the focal school. The notation F in the equations below is used to indicate the latent class in which the focal school is located, i.e., the focal group. The remaining $K-1$ latent classes are each separately used as reference groups for this DIF analysis. If $K \geq 3$, then the DIF analysis is a multi-group DIF analysis.

The school-level DIF measures are based on differences in item difficulties between the focal group and each of the $K-1$ other latent classes separately defined as follows for binary and for polytomous items, respectively:

$$DIFM = \hat{\beta}_{Tigk} - \hat{\beta}_{TigK}, \tag{3.37}$$

and

$$DIFM = \hat{\delta}_{Tigk} - \hat{\delta}_{TigK}, \tag{3.38}$$

where $\hat{\beta}_{Tigk}$ and $\hat{\delta}_{Tigk}$ are estimated item difficulties transformed.

Let the group $k = 1$ be the focal group for the focal school and the remaining $K-1$ groups (i.e., $k = 2, \ldots, K$) be the reference groups for that school. For the studied item i and for student-level latent class g, the following comparisons are of interest:

$$\mathrm{H}_0 : \beta_{Tigk} - \beta_{TigK} = 0, \tag{3.39}$$

and

$$\mathrm{H}_0 : \delta_{Tigk} - \delta_{TigK} = 0 \ , \tag{3.40}$$

indicating that there is no difference in item difficulties between two school-level groups.

DIF Detection Procedure. Since Bayesian estimation is used to estimate model parameters, then a Bayesian method for estimation of DIF needs to be used as well. In this regard, the difference between each pair of item difficulties can be tested with a credibility interval. Box and Tiao (1973) describe this interval as the highest posterior density (HPD) interval. The HPD interval containing 95 percent of the probability under the posterior is referred to as the 95 percent highest posterior density (HPD) interval. Assuming a nominal alpha of .05 for rejection of the null hypothesis in which a parameter value or a function of parameter values is 0, the HPD interval can be used to test, if the value differs significantly from zero (Box & Tiao, 1973). In this study, if the HPD interval of the DIF measure for each item does not include 0, the item is considered as DIF item (Samuelsen, 2005).

3.3 Simulation design

Simulation of Mixtures. Rost (1990) suggested the primary diagnostic potential of the MRM (i.e., without school-level mixtures) is in its use for accounting for qualitative differences among examinees. De Boeck et al. (2005) describe two types of qualitative differences between manifest categories. The qualitative differences are "simple" when the location of item difficulties has shifted relative to the location of the other item difficulties. The qualitative differences are "complex" when there does not appear to be a discernible pattern to the locations of the item difficulties among the latent classes. Likewise, two types of qualitative differences at the school-level mixture are possible: simple and complex qualitative differences. In this study, complex qualitative differences were simulated among both school-level and student-level latent classes.

Mixtures were simulated to represent different item difficulty patterns for each class at the student- and school-levels. In a mathematics test used in the PSAT/NMSQT, there are four math concepts in mathematics section (i.e., number of operations, algebra and functions, geometry and measurement, and data analysis, statistics, and probability). Likewise, there are three kinds of problems of writing skills questions in the writing test in the PSAT/NMSQT (i.e., improving sentences, identifying sentence errors, and improving paragraphs). Four knowledge levels were assumed for generating item difficulty parameters are generated. It was assumed that each concept was measured by 10 items. Further, each school-level latent class was assumed to have its own student-level class-specific item difficulty. Item difficulty parameters were set as shown in Tables 3.1 and 3.2. In this study, school-level DIF was generated by increasing and decreasing the β parameters by different magnitudes. Two levels of the percent of DIF items on the test were simulated: Four items (i.e., 10 percent items in the test) and 12 items (i.e., 30 percent of the items in the test) were generated as DIF items at the school-level. Ability was simulated as $\theta \sim$ Normal $(0, 1)$ for each latent class (i.e., g and k) for these item difficulties, resulting in multilevel mixture structures.

Simulation Conditions. A simulation study was done to describe the detection of DIF for the MMixIRTM under typical testing conditions. The following conditions were examined in the simulation study: test length, effect size of DIF, numbers of items with DIF, numbers of latent classes at the student- and school-level, student and school sizes, proportion of mixtures, modeling of probabilities of mixtures, and strength of relationship between covariates and mixtures. Strength of relationship between covariates and mixtures condition was examined only for the multinomial regression modeling on the probabilities of mixtures.

There were a total of 32 conditions simulated for the use of priors: 1 test length $\times$ 1 magnitude of the differences in item difficulties $\times$ 2 levels of the percent of items with DIF $\times$ 2 latent classes at the student- and school-levels $\times$ 2 numbers of students for each school

$\times$ 1 total number of examinees $\times$ 1 proportion of mixtures $\times$ 8 different models fit to each data set.

There were a total of 48 conditions simulated to investigate the use of a multinomial logistic regression model: 1 test length $\times$ 1 magnitude of the differences in item difficulties $\times$ 2 levels of the percent of items with DIF $\times$ 1 class at the student- and school-level $\times$ 2 numbers of students for each school $\times$ 1 total number of examinees $\times$ 1 proportion of mixtures $\times$ 2 strength of relationship between covariates and mixtures $\times$ 6 model selection (2 models are overlapped with models with the priors).

Thus, 80 conditions were simulated with five replications for each condition, yielding a total of 400 simulated data sets analyzed. The simulation conditions are described below.

Test length. A recovery study for the MMixIRTM for a fixed length test of 40-items was used. The length of 40 items was used since this was similar to the number of items on the sections of the PSAT/NMSQT. That is, there are 39 items on the Writing Skills section and 38 items on the Mathematics section (28 multiple-choice items and 10 gridded items).

Magnitudes of the differences in item difficulties. Differences among latent classes are larger than typically observed among manifest groups (Samuelsen, 2005). This is a function of the clustering which is done by the MixIRTMs. These models function to detect groups that are homogeneous with respect to the latent ability and also to increase differences on the latent ability among different latent classes. This result was also observed in studies of DIF based on analysis of latent groups by Cohen and Bolt (2005) and by Cohen et al. (2005).

The size of DIF was simulated as the difference between the item difficulty parameters among the classes. DIF was introduced into items by increasing and decreasing the β parameters by 0.4, 0.6, 0.8, 1, and 1.2 magnitudes, which reflect the "complex" qualitative difference at the school level. The strategy of increasing and decreasing the β parameters by magnitudes are used in the Smit et al. (1999) and Samuelsen (2005) at the student-level.

Smit et al. (1999) used differences of 1 and 2; Samuelsen (2005) used differences of 0.4, 0.8, and 1.2.

Numbers of Items with DIF. 10 and 30 percent of DIF items at the school-level will be used. The percentages of DIF items simulated were based on previous research which suggests that it is not uncommon to observe between 10% and 30% of the items to function differentially on many tests (Hambleton & Rogers, 1989; Raju, Bode, & Larsen, 1989). Such percentages are larger than would typically be observed in operational forms of well-constructed tests, but it also was expected that more DIF items would be observed with a latent classes approach than with a manifest groups approach (Samuelsen, 2005).

Numbers of Latent Classes at the Student- and School-Level. The simulation study examined the number of students and schools, and the number of latent classes at the student-level and at the school-level. These conditions are described in Table 3.3 .

In Table 3.3, I indicates the case when $g < k$, II indicates the case when $g = k$, and III indicates the case $g > k$. For $g = 1$ and $k = 1$, the model is actually Rasch model or a multilevel IRT model rather than a MMixIRTM. In addition, when $g = 2$ or $g = 3$ and $k = 1$, the model is actually a MixIRTM. The multilevel IRT model and MixIRTM in Table 3.3 will be considered in this study but will be excluded for the recovery study portion. Model recovery of a MixIRTM with fully Bayesian estimation is done in the Cho et al. (2006) and in Li et al. (2006).

In modeling of a MMixIRTM, the combination of $g = 1$ with $k \geq 1$ is not meaningful in the current study, because the number of school-level latent classes, k, is composed of different proportions of the g. In addition, it is less likely to have a condition of $g = 2$ and $k = 3$, as there is a smaller likelihood that two different latent classes at the student-level will be capable of forming three different proportions of mixtures at the school level. In this study, $g = 2$ and $k = 2$ were simulated.

Student and School Sample Sizes. As mentioned earlier, Cho et al. (2006) found that some of the MCMC chains had not converged by the time they had reached 100,000

iterations. This occurred for all the 10-item conditions for each sample size and for the 30-item and 50-item conditions with 360 examinees as well. These results suggest that the data for these conditions were not sufficiently informative to obtain stable parameter estimates for these conditions. The results also suggest that test lengths for the MRM should be at least 30 items and sample sizes should be at least 1,000 examinees. Finally, Cho et al. found that, when the sample size was small, DIF was detected, if the overlap between the latent classes and manifest groups was greater than 80 percent and the magnitude of the DIF was large.

The proposed MMixIRTM is a complex model in that it assumes a multilevel latent class structure. In addition, a hierarchical data structure is incorporated into the model so it is appropriate for use with school-level data. The effect of the number of students in each school was considered in this study as it is likely of more interest than the total number of examinees. In addition, it was of interest how many students and schools were needed in each latent class for recovery of parameters. This was be investigated by considering the proportions of each class (g and k) using the different conditions shown in Table 3.4. The total sample size as well as the number of students for each school were considered in determining how many students and schools were included in each class.

With respect to student- and school-sample sizes, we focused on MMixIRTM behavior with respect to the ratio of the number of examinees to the number of latent classes. The efficiency and power of multilevel tests rests on pooled data across the units comprising two or more levels. This typically implies large datasets. Simulation studies by Kreft (1996) found adequate statistical power with 30 groups of 30 observations each, 60 groups with 25 observations each, and 150 groups with 5 observations each.

Currently, the SOAS report from The College Board is provided to each school that has 25 or more students participating in the PSAT/NMSQT program. 25, therefore, was the number of students chosen as the small student sample size and 100 as the large size. One fixed number of examinees, 8,000, was used for the simulations. This yielded 320 schools for

the small student sample size and 80 schools for the large student sample size. The two ratios for these sample sizes were .078 and 1.25, respectively. These two sample sizes of the number of students for each school and the fixed number of the examinees provide two different sizes of the number of schools for each latent class.

Proportion of mixtures. Equal proportions of membership in school-level latent class were simulated in this study. In the two-group simulations, the proportions were both .50. Table 5 describes the proportion of each latent class for each combination of student-level $\times$ school-level mixture for the $g = 2$ and $k = 2$ condition. For $K = 1$, the first student-level class (i.e., $G = 1$) was simulated to have a dominant group with 92 percent of the examinees in the first school-level class. For $K = 2$, the second student-level class (i.e., $G = 2$) was simulated as the dominant group with 68 percent of the cases. This reflected a less dominant group in the second school-level latent class.

Modeling of Probabilities of Mixtures. Preliminary checking of priors and hyper-parameter combinations suggested the Dirichlet distribution with Gamma sampling for both student-level and school-level probabilities of mixtures should yield good recovery of school-level group membership. In addition, probabilities of mixtures from a multinomial logistic regression model also were used for both student- and school-level as another condition.

Strength of Relationship Between Covariates and Recovery of Mixtures. This condition applies only to the use of a multinomial logistic regression model on mixture probabilities. Previous research suggests that the relationship between a covariate and the mixture is a potentially important factor in the accuracy of detection of latent classes (Smit et al., 1999; Cho et al., 2006). Results from Cho et al. (2006) suggest the use of a prior with a covariate improved recovery of item difficulties and the probabilities of latent classes for three and four-class models when overlap was 80 percent or more and no guessing was simulated. In addition, Samuelsen (2005) has shown that, as the amount of overlap between manifest groups and latent classes decreased, so did the power to correctly detect DIF items.

Two levels of overlap were simulated between manifest and latent groups: 60 percent overlap, and 100 percent overlap for each level of latent class. This condition will not be fully crossed, however, for each level of number of latent classes. One hundred percent overlap referred to a condition in which one latent class was composed entirely of examinees from a single manifest group and the other overlap conditions refers to situations in which the latent class was increasingly more heterogeneous with respect to the manifest groups (Samuelsen, 2005). Sixty percent overlap means 40 percent of the people in the manifest groups behaved like those in the other manifest group. The resulting numbers of examinees within the simulated classes are shown in Table 3.5.

Model selection. Eight different numbers of mixtures were fitted to each number of student- and school-level mixture. Frühwirth-Schnatter (2006) noted that label switching is unavoidable for a model that is overfitting the number of mixtures because one or more than probabilities of mixtures can be 0, or there are the possibilities that model parameters are equal. Each combination of numbers of mixtures was monitored if the model is identified or not first before AIC and BIC values are obtained.

Models were checked for $G = 1, \ldots, 4$ at the student-level and $K = 1, \ldots, 4$ at the school-level. However, three combinations, $G = 1$ and $K = 2$, $G = 1$ and $K = 3$, and $G = 1$ and $K = 4$, were not simulated as those two cases do not make sense in the present context. That is, if there is no student-level mixture (i.e., $G = 1$), a school-level mixture will not exist.

Recovery Evaluation. The accuracy of parameter estimation is dependent on having examinees classified correctly, and classification accuracy is dependent on accurate parameter estimation (Bolt, Cohen, & Wollack, 2002). Thus, only conditions that for recovery of item difficulties and latent class memberships for the correct number of latent classes were considered .

For the recovery of item difficulty, the root mean square error (RMSE), bias (Bias), and Pearson correlations (Corr) were computed across replications with the following equations:

$$RMSE(b) = \sqrt{\frac{1}{IGK}\sum_{i=1}^{I}\sum_{g=1}^{G}\sum_{k=1}^{K}(\hat{b_{igk}} - b_{igk})^2}, \tag{3.41}$$

$$Bias(b) = \frac{1}{IGK}\sum_{i=1}^{I}\sum_{g=1}^{G}\sum_{k=1}^{K}[E(\hat{b_{igk}}) - b_{igk})], \tag{3.42}$$

and

$$Corr = \frac{Cov(\hat{b_{igk}}, b_{igk})}{\sigma_{\hat{b_{igk}}}, \sigma_{b_{igk}}}. \tag{3.43}$$

Recovery of latent group membership was assessed by comparing the estimated group membership with the generating parameters. This was done at the mode of the sampled group memberships following burn-in.

Linking of Scales for RMSE and Bias Analyses. To compare estimates of model parameters with the generating parameters using the RMSE and bias statistics, it was necessary first to place the estimates and the generating parameters on the same scale. Pearson correlations do not require both variables to be placed on the same scale. In this study, the estimates were placed on the scale of the generating parameters. The following adjustment was needed to calculate RMSE and bias statistics:

$$\beta_T^* = \beta_T - (\bar{\beta_T} - \bar{\beta_B}), \tag{3.44}$$

T represents target scale, that is, the estimates scale in the simulation study and B does base scale, that is, the true parameter scale in the simulation study.

Table 3.1: Generating Parameters for Mixture Model Simulations: Patterns with 30% Complex Qualitative Difference at the School-Level

			Generating Item Parameters			
Item	DIF Magnitude at $G = 1$	DIF Magnitude at $G = 2$	$G = 1, K = 1$	$G = 2, K = 1$	$G = 1, K = 2$	$G = 2, K = 2$
1	-0.4	0.4	-2.00	0.00	-2.40	0.40
2	-1.2	1.2	-2.00	0.00	-3.20	1.20
3	-0.8	0.8	-1.75	0.00	-2.55	0.80
4	-1.0	1.0	-1.75	0.25	-2.75	1.25
5	-0.6	0.6	-1.50	0.25	-2.10	0.85
6	-1.2	1.2	-1.50	0.25	-2.70	1.45
7	-0.4	0.4	-1.25	0.25	-1.65	0.65
8	-1.2	1.2	-1.25	0.50	-2.45	1.70
9	-0.4	0.4	-1.00	0.50	-1.80	1.30
10	-0.8	0.8	-1.00	0.50	-2.00	1.50
11	-1.0	1.0	-0.50	1.00	-1.10	1.60
12	-0.6	0.6	-0.50	1.00	-1.70	2.20
13	-1.2	1.2	-0.50	1.00	-0.50	1.00
14	0.0	0.0	-0.25	1.00	-0.25	1.00
15	0.0	0.0	-0.25	1.25	-0.25	1.25
16	0.0	0.0	-0.25	1.25	-0.25	1.25
17	0.0	0.0	-0.25	1.25	-0.25	1.25
18	0.0	0.0	0.50	1.25	0.50	1.25
19	0.0	0.0	0.50	1.50	0.50	1.50
20	0.0	0.0	0.50	1.50	0.50	1.50
21	0.0	0.0	0.00	-0.50	0.00	-0.50
22	0.0	0.0	0.00	-0.50	0.00	-0.50
23	0.0	0.0	0.00	-0.50	0.00	-0.50
24	0.0	0.0	0.25	-0.25	0.25	-0.25
25	0.0	0.0	0.25	-0.25	0.25	-0.25
26	0.0	0.0	0.25	-0.25	0.25	-0.25
27	0.0	0.0	0.25	-0.25	0.25	-0.25
28	0.0	0.0	0.50	0.50	0.50	0.50
29	0.0	0.0	0.50	0.50	0.50	0.50
30	0.0	0.0	0.50	0.50	0.50	0.50
31	0.0	0.0	1.00	-2.00	1.00	-2.00
32	0.0	0.0	1.00	-2.00	1.00	-2.00
33	0.0	0.0	1.00	-1.75	1.00	-1.75
34	0.0	0.0	1.00	-1.75	1.00	-1.75
35	0.0	0.0	1.25	-1.50	1.25	-1.50
36	0.0	0.0	1.25	-1.50	1.25	-1.50
37	0.0	0.0	1.25	-1.25	1.25	-1.25
38	0.0	0.0	1.25	-1.25	1.25	-1.25
39	0.0	0.0	1.50	-1.00	1.50	-1.00
40	0.0	0.0	1.50	-1.00	1.50	-1.00

Table 3.2: Generating Parameters for Mixture Model Simulations: Patterns with 10% Complex Qualitative Difference at the School-Level

			Generating Item Parameters			
Item	DIF Magnitude at $G = 1$	DIF Magnitude at $G = 2$	$G = 1, K = 1$	$G = 2, K = 1$	$G = 1, K = 2$	$G = 2, K = 2$
1	-0.4	0.4	-2.00	0.00	-2.40	0.40
2	-1.2	1.2	-2.00	0.00	-3.20	1.20
3	-0.8	0.8	-1.75	0.00	-2.55	0.80
4	-1.0	1.0	-1.75	0.25	-2.75	1.25
5	0.0	0.0	-1.50	0.25	-1.50	0.25
6	0.0	0.0	-1.50	0.25	-1.50	0.25
7	0.0	0.0	-1.25	0.25	-1.25	0.25
8	0.0	0.0	-1.25	0.50	-1.25	0.50
9	0.0	0.0	-1.00	0.50	-1.00	0.50
10	0.0	0.0	-1.00	0.50	-1.00	0.50
11	0.0	0.0	-0.50	1.00	-0.50	1.00
12	0.0	0.0	-0.50	1.00	-0.50	1.00
13	0.0	0.0	-0.50	1.00	-0.50	1.00
14	0.0	0.0	-0.25	1.00	-0.25	1.00
15	0.0	0.0	-0.25	1.25	-0.25	1.25
16	0.0	0.0	-0.25	1.25	-0.25	1.25
17	0.0	0.0	-0.25	1.25	-0.25	1.25
18	0.0	0.0	0.50	1.25	0.50	1.25
19	0.0	0.0	0.50	1.50	0.50	1.50
20	0.0	0.0	0.50	1.50	0.50	1.50
21	0.0	0.0	0.00	-0.50	0.00	-0.50
22	0.0	0.0	0.00	-0.50	0.00	-0.50
23	0.0	0.0	0.00	-0.50	0.00	-0.50
24	0.0	0.0	0.25	-0.25	0.25	-0.25
25	0.0	0.0	0.25	-0.25	0.25	-0.25
26	0.0	0.0	0.25	-0.25	0.25	-0.25
27	0.0	0.0	0.25	-0.25	0.25	-0.25
28	0.0	0.0	0.50	0.50	0.50	0.50
29	0.0	0.0	0.50	0.50	0.50	0.50
30	0.0	0.0	0.50	0.50	0.50	0.50
31	0.0	0.0	1.00	-2.00	1.00	-2.00
32	0.0	0.0	1.00	-2.00	1.00	-2.00
33	0.0	0.0	1.00	-1.75	1.00	-1.75
34	0.0	0.0	1.00	-1.75	1.00	-1.75
35	0.0	0.0	1.25	-1.50	1.25	-1.50
36	0.0	0.0	1.25	-1.50	1.25	-1.50
37	0.0	0.0	1.25	-1.25	1.25	-1.25
38	0.0	0.0	1.25	-1.25	1.25	-1.25
39	0.0	0.0	1.50	-1.00	1.50	-1.00
40	0.0	0.0	1.50	-1.00	1.50	-1.00

Table 3.3: Model Selection for the Number of Latent Classes

	$K = 1$	$K = 2$	$K = 3$
$G = 1$	II (MulIRT Model)	I (Unavailable)	I (Unavailable)
$G = 2$	III(MixIRT Model)	II	I
$G = 3$	III(MixIRT Model)	III	II

Table 3.4: Proportions Simulated in Each Latent Class

	$P(K = 1)$=.5	$P(K = 2)$=.5
$P(G = 1)$	.92	.32
$P(G = 2)$	.08	.68

Table 3.5: Numbers of Examinees Within Latent Classes

Overlap		LC1	LC2
100	M1	4000	0
	M2	0	4000
60	M1	2400	600
	M2	600	2400

Chapter 4

Results

4.1 Result of simulation study

4.1.1 Detection of Label Switching

As described earlier, there are three possible types of label switching in MMixIRTM. Given that the estimated marginal posterior densities for π_k in the MMixIRTM were unimodal, it was appropriate to conclude that there was no label-switching within the MCMC chains for both models with and without covariates. That is, the first type of label-switching was not observed, and a necessary condition for identification was satisfied.

However, we did find the second type of label switching (i.e., label switching across chains) for a few of the simulation conditions for both the models with and without covariates. By comparing parameter estimates with the generating parameters, it was possible to determine which labels should be applied to each of the latent classes. The third type of label switching was not detected since labelling of student-level group membership was consistent across school-level mixtures for models with and without covariates.

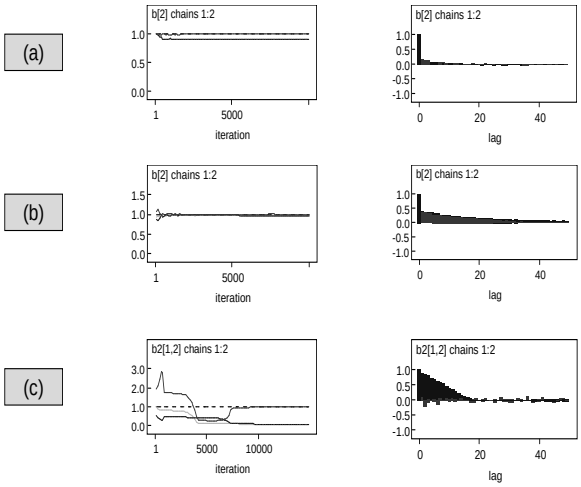
4.1.2 Monitoring Convergence

In this study, evidence was examined from four convergence methods, the Gelman and Rubin (1992) method as implemented in WinBUGS, and the Geweke (1992) method, and the Raftery and Lewis (1992) method as implemented in the computer program Bayesian Output Analysis (Smith, 2004). In addition, evidence was examined from autocorrelation plots from WinBUGS.

The Rasch model, multilevel Rasch model, the MRM, and the MMixIRTM all were fitted for each dataset generated as a MMixIRTM with $G = 2$ and $K = 2$ (as described in the section on model selection). The Rasch model and multilevel Rasch model converged within 100 iterations for all simulated conditions. The two-group solution for the MRM converged within 6,000 iterations for all simulated conditions. The three-group solution for the MRM did not converge for most item difficulty parameters except for the condition with 25 students per school and 30 percent DIF items. Given this result, it seems likely that the three-group solution was not a reasonable approximation to data generated for two groups.

Figure 4.1 presents plots of the Gelman and Rubin statistics and autocorrelations for a difficulty parameter of a single item as an example to show what the convergence looked like for the simulation condition 10 % DIF items and 25-student/320-school. In Figure 4.1, the left column presents the plot of the Gelman and Rubin statistics and the right column presents the plot of autocorrelations. The Gelman and Rubin statistic approaches 1 quickly and the level of the autocorrelations was low, which indicates convergence. Similar patterns were observed for the other parameters and conditions.

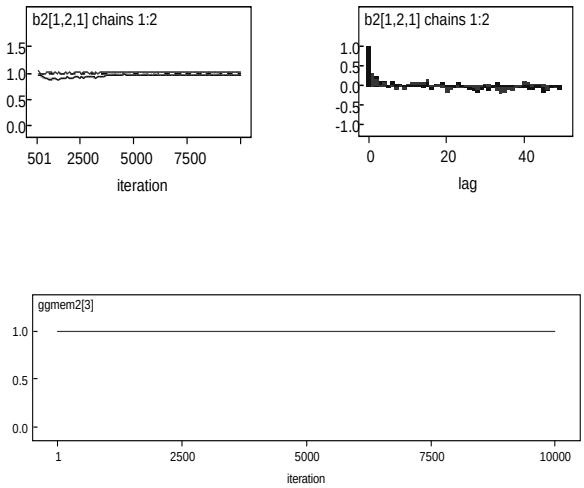
Figure 4.1: Selected Plots of Gelman and Rubin Statistic and Autocorrelation: (a) Rasch Model, (b) Multilevel Rasch Model, and (c) MRM: 2-Group Solution



For the MMixIRTM two-group solution at the student-level, two- and three-group solutions at the school-level converged within 5,000 iterations, although there were relatively high autocorrelations for some item parameter estimates. To deal with this, the chains were thinned by sampling every 40 iterations. The result was the autocorrelations decreased. The top of Figure 4.1 shows the plots of the Gelman and Rubin statistics and autocorrelations for a difficulty parameter of one item as an example to represent the convergence for the simulation condition 10 % DIF items and 25-student/320-school for the MMixIRTM. For other models, Gelman and Rubin statistic also approached 1 quickly and the level of autocorrelations was low after thinning to sample every 40th iteration. These results indicate convergence. Other parameters and conditions had similar patterns to those shown in Figure 4.1.

The chain for estimating group membership at the school-level became stuck at initial values for the three-group student-level simulation for the two- and three-group school-level solutions as shown in the lower part of Figure 4.2. This may have occurred because a three-group solution for the MRM may actually be overfitting given the generated data.

Figure 4.2: Figures on MMixIRTM



For this study, a 7,000 iteration burn-in and an 8,000 iteration post-burn-in were used across all conditions. Since thinning was set at 40 for MMixIRTM, this meant that 32,000 iterations were required after burn-in.

4.1.3 Simulation Results

Model selection

The recovery of the generating model was assessed using a model selection approach in which two information-based indices, AIC and BIC, were used to determine model fit. The two indices were calculated inside the MCMC algorithm and compared as averages over the post-burn-in iterations (Congdon, 2003). The usual approach in mixture modeling, and the one assumed in this study, was to fit a range of appropriate candidate mixture models each with a different number of latent classes, and then to consider the selection of a model according to some theoretical rationale using one or more statistical criteria to guide the choice. Results from Li et al. (2006) suggested that the BIC was more accurate than the AIC for model selection with MixIRTMs. AIC tended to select the more complex model (McLachlan & Peel, 2000; Cho et al., 2006; Li et al., 2006). In this study, therefore, AIC is shown (because it is a commonly reported index), but correct models were determined based only on the value of BIC. Tables 4.1, 4.2, 4.3, 4.4, 4.5, and 4.6 show the probabilities that a model had the minimum AIC and BIC across iterations after burn-in. The probabilities are averaged across replications.

Table 4.1: Model Selection Result: AIC and BIC for 10 Percent DIF Items with a Prior on the Probabilities of Mixtures

Number of Mixtures	Model	25 students/320 schools BIC	25 students/320 schools AIC	100 students/80 schools BIC	100 students/80 schools AIC
$G = 1$ $K = 1$	Rasch Model	0	0	0	0
$G = 1$ $K = 1$	Multilevel Rasch Model	0	0	0	0
$G = 2$ $K = 1$	Mixture IRT Model	**0.93**	0.13	**0.88**	0.14
$G = 2$ $K = 2$	Multilevel mixture IRT Model	0.07	**0.66**	0.12	**0.72**
$G = 2$ $K = 3$	Multilevel mixture IRT Model	0	0.21	0	0.13

Table 4.2: Model Selection Result: AIC and BIC for 30 Percent DIF Items with a Prior on the Probabilities of Mixtures

Number of Mixtures	Model	25 students/320 schools BIC	25 students/320 schools AIC	100 students/80 schools BIC	100 students/80 schools AIC
$G = 1$ $K = 1$	Rasch Model	0	0	0	0
$G = 1$ $K = 1$	Multilevel Rasch Model	0	0	0	0
$G = 2$ $K = 1$	Mixture IRT Model	0.11	0	0.08	0
$G = 2$ $K = 2$	Multilevel mixture IRT Model	**0.89**	**0.74**	**0.92**	**1**
$G = 2$ $K = 3$	Multilevel mixture IRT Model	0	0.26	0	0

Table 4.3: Model Selection Result: AIC and BIC for 10 Percent DIF Items with a Multinomial Regression Model with a 60 Percent Overlap on the Probabilities of Mixtures

Number of Mixtures	Model	25 students/320 schools BIC	AIC	100 students/80 schools BIC	AIC
$G = 1$ $K = 1$	Rasch Model	0	0	0	0
$G = 1$ $K = 1$	Multilevel Rasch Model	0	0	0	0
$G = 2$ $K = 1$	Mixture IRT Model	**0.80**	0.22	**0.72**	0.22
$G = 2$ $K = 2$	Multilevel mixture IRT Model	0.20	**0.59**	0.11	0.32
$G = 2$ $K = 3$	Multilevel mixture IRT Model	0	0.19	0.17	**0.46**

Table 4.4: Model Selection Result: AIC and BIC for 10 Percent DIF Items with a Multinomial Regression Model with a 100 Percent Overlap on the Probabilities of Mixtures

Number of Mixtures	Model	25 students/320 schools BIC	AIC	100 students/80 schools BIC	AIC
$G = 1$ $K = 1$	Rasch Model	0	0	0	0.00
$G = 1$ $K = 1$	Multilevel Rasch Model	0	0	0	0.00
$G = 2$ $K = 1$	Mixture IRT Model	**0.82**	0.02	**1**	0.11
$G = 2$ $K = 2$	Multilevel mixture IRT Model	0.18	**0.78**	0	0.36
$G = 2$ $K = 3$	Multilevel mixture IRT Model	0	0.20	0	**0.53**

As shown in Tables 4.1, 4.2, and 4.3 for the 10-percent DIF condition, the correct number of mixtures (i.e., $G = 2$ and $K = 2$) was not selected by using BIC. In the 30-percent DIF condition, however, BIC accurately selected the correct model for both 25 and 100 students per school. Consistent with previous research, AIC tended to select the more complex model. BIC selected the MRM with a two-group solution for the 10-percent DIF condition. This result implied that more than 10 percent DIF was needed on a test at the school-level to detect the correct number of latent classes at student- and school-levels.

Table 4.5: Model Selection Result: AIC and BIC for 30 Percent DIF Items with a Multinomial Regression Model with a 60 Percent Overlap on the Probabilities of Mixtures

Number of Mixtures	Model	25 students/320 schools BIC	AIC	100 students/80 schools BIC	AIC
$G = 1$ $K = 1$	Rasch Model	0	0	0	0
$G = 1$ $K = 1$	Multilevel Rasch Model	0	0	0	0
$G = 2$ $K = 1$	Mixture IRT Model	0.18	0	0.07	0
$G = 2$ $K = 2$	Multilevel mixture IRT Model	**0.82**	**0.75**	**0.93**	**0.83**
$G = 2$ $K = 3$	Multilevel mixture IRT Model	0	0.25	0	0.17

Recovery of Generating Parameters

A recovery analysis was done to determine whether the MCMC algorithm programmed in WinBUGS accurately recovered the generating parameters under the conditions simulated. The accuracy of parameter estimation is dependent on having examinees classified correctly and classification accuracy is likewise dependent on accurate parameter estimation (Bolt, Cohen, & Wollack, 2001). Thus, the recovery study was conducted only for conditions with the correct numbers of latent classes. One replication of a condition was chosen to indicate recovery of item difficulty parameters and transformed item difficulty estimates (see Table 4.9).

The recovery results for the 30-percent DIF condition is shown in Tables 4.7 and 4.8.

The recovery of group membership for both student- and school-level mixtures was good for both 25 students/320 schools and 100 students/80 schools. With respect to the correctly selected number of mixtures and identified model, the accuracy of detection of student-level group membership was $98.5\% - 99.6\%$ and the accuracy of detection of school-level group membership was 100%. RMSE of item difficulty was $.094 - .115$, bias was $-.001 - .033$,

Table 4.6: Model Selection Result: AIC and BIC for 30 Percent DIF Items with a Multinomial Regression Model with a 100 Percent Overlap on the Probabilities of Mixtures

Number of Mixtures	Model	25 students/320 schools BIC	25 students/320 schools AIC	100 students/80 schools BIC	100 students/80 schools AIC
$G = 1$ $K = 1$	Rasch Model	0	0	0	0
$G = 1$ $K = 1$	Multilevel Rasch Model	0	0	0	0
$G = 2$ $K = 1$	Mixture IRT Model	0.05	0	0.01	0
$G = 2$ $K = 2$	Multilevel mixture IRT Model	**0.95**	**0.82**	**0.99**	**0.89**
$G = 2$ $K = 3$	Multilevel mixture IRT Model	0	0.18	0	0.11

Table 4.7: Model Parameter Recovery with 30 Percent DIF Items: Item Difficulty Recovery

		25 Students/320 Schools			100 Students/80 Schools		
		RMSE	Bias	Correlation	RMSE	Bias	Correlation
Prior		0.096	0.033	.997	0.101	0.000	.997
Covariate	60 percent	0.097	-0.001	.997	0.100	0.000	.997
	100 percent	0.094	0.000	.997	0.115	0.001	.997

Pearson correlations were .997 for all 30% DIF items conditions. This result indicates that the item difficulty parameter are recovered well.

4.2 Empirical illustration: Mathematics Section

In this section, the proposed model, MMixIRTM, is illustrated with the PSAT/NMSQT Mathematics section. The sample was selected with respect to the definition of the comparable group, which is defined as students in a group in the 10th or 11th grade for the current school year, who have scores between 20 and 80 inclusive, excluding non-standard students.

Table 4.8: Model Parameter Recovery with 30 Percent DIF Items: Group Membership Recovery

		25 Students/320 Schools		100 Students/80 Schools	
		Student-Level	School-Level	Student-Level	School-Level
Prior		98.7	100	98.6	100
Covariate	60 percent	98.7	100	98.6	100
	100 percent	99.4	100	99.9	100

In addition, the schools having more than 25 students were considered. In doing so, we show how heterogeneous the comparable group is with model-based clustering. The data includes 987 schools and $39,614$ students. An approximately 20% random sub-sample of 206 schools and $8,362$ students is used for the empirical illustration.

The first part in this section describes the data for the MMixIRTM preliminary analysis. Next, the results for the MMixIRTM and STD P-DIF analyses are shown and then compared with regard to the detected DIF items at the school-level.

4.2.1 Data

Before using the model it was necessary to investigate the adequacy and plausibility of the model. In this regard, the multilevel structure was investigated. In addition, covariates associated with students' achievement are reviewed below.

Investigation of Hierarchical Structure. Before fitting a multilevel model to the data, it was first necessary to determine whether a hierarchical structure was present in the data. The standard technique is to use intra-class correlations (ICC, Raudenbush & Bryk, 2002) for both total score and latent score. Two ICC values are shown with a multilevel IRT model using "latent item score" and with a multilevel analysis using "observed sum score" (i.e., the sum of the number of correct answers).

Fox and Glas (2001) note that the use of latent as opposed to observed scores as dependent variables in a multilevel model provides the potential for separation of the effects of item difficulty and ability, and also enables the modeling of response measurement error. Unlike observed scores, latent scores are also invariant over items meaning it is possible to use different forms of a test and estimating parameters for the items and the multilevel model simultaneously.

For the multilevel IRT model, estimated student-level variance was .946, and estimated school-level variance was .385. Both were statistically significant. The ICC for those two estimates were .289, which indicated 28.9% of the student-level variance was explained at the school level.

The dependent variable in a multilevel analysis is the sum of the number of correct answers. This variable was standardized to be normal and no predictors were used to obtain the ICC with varying intercept. A linear mixed-effects model was fitted using the *lmer* function (Bates & Debroy, 2004) written in *R* (R Development Core Team 2007). The estimated student-level variance was .794, and the estimated school-level variance was .223. The ICC for these two estimates was .219, which indicates 21.9% of the student-level variance was explained at the school level.

Demographic information. Since gender and ethnicity have been shown to be associated with student math achievement (Maple & Stage,1991; Tate, 1997) both variables were considered as a student-level covariates. School-level variables were grouped into three categories using information available from the PSAT/NMSQT data file as follows: (1) the structural feature of the school including school location and school enrollment size, (2) school-level SES variables including Title I schoolwide program, household income codes, and poverty-level code, and (3) minority percentage for Caucasian, African-American, Native-American, Asian, and Hispanic. In this study, the structural feature of the school and school-level SES variables were considered for analysis.

4.2.2 MMixIRTM Result

Fitting MMixIRTM: Statistical Analysis

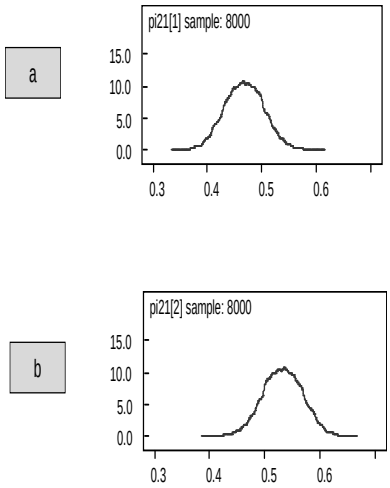
Model Identification Checking and Label Switching Analysis. As mentioned earlier, there are three possible types of label switching that can arise in a MMixIRTM. That is, label switching within a chain as Type I, label switching across chains as Type II. And Type III label switching, a variant of Type II switching, in which student-level latent classes switch within a school-level latent class. These three types of label switching were considered in the empirical data analysis.

Since the pattern of three types of label switching was similar for both with and without covariates models, the label switching analysis for only the model without covariates is reported below. The necessary condition for identification of a MMixIRTM is that the school-level classes should be identified. Since there was no multimodality in the density of $\hat{\pi_k}$ (as can be seen in Figure 4.3), it was concluded that a necessary condition for identification was satisfied and that there was no label switching within a chain (i.e., Type I label switching did not occur). In addition, there was also no multimodality in the density of $\hat{\pi_{g|k}}$.

However, we did find label switching across chains (i.e., Type II label switching was observed). This was detected in convergence checking with more than two chains and was evident as the patterns of the means of ability and mixture proportions were different across chains. In addition, we investigated the possibility of the Type III label switching by examining the patterns of class-specific item difficulties at the student-level across school-level mixtures. Since patterns of relative difficulty for certain items appear to be similar in a consistent manner across school-level mixtures, we concluded that there was no Type III label switching.

Model Selection. Results from Li et al. (2006) suggested that the BIC was more accurate than the AIC for model selection with MixIRTMs. In this study, therefore, both AIC and BIC were computed for the comparisons. Model selection, however, was determined based only on the value of BIC. Tables 4.10 and 4.11 present the averaged AIC and BIC values

Figure 4.3: Marginalized Density Function for the Probability of Mixtures at the School-Level: (a) For School-Level Class 1, (b) For School-Level Class 2



across iterations after burn-in and the probabilities that a model had the minimum AIC and the minimum BIC over the MCMC chain after burn-in.

Based on the BIC values shown in Table 4.10, $G = 4$ and $K = 2$ were chosen based for the model without covariates. BIC values in Table 4.11 show that $G = 4$ and $K = 2$ were chosen for the model with covariates.

Model Convergence Checking. We examined convergence using three methods, the Gelman and Rubin (1992) method as implemented in WinBUGS, and the Geweke (1992) method and the Raftery and Lewis (1992) method as implemented in the computer program Bayesian Output Analysis (Smith, 2004) for both without covariate and with covariate models. The Gelman and Rubin (1992) statistic was run for two chains using widely disparate initial values for each of the parameters (e.g., the initial values for the difficulty parameters, b, were set at -2 and 2, respectively, for the two chains, for the deviation of

ability, σ, the initial values were 1 and 2, respectively, for the two chains). Since the second-type of label switching did occur for with covariate model, we got Gelman-Rubin statistic after we matched the labelling of mixtures across chains by looking at the pattern of the mean of ability and mixture proportions. The Geweke (1992) and the Raftery and Lewis (1992) methods were run for a single chain.

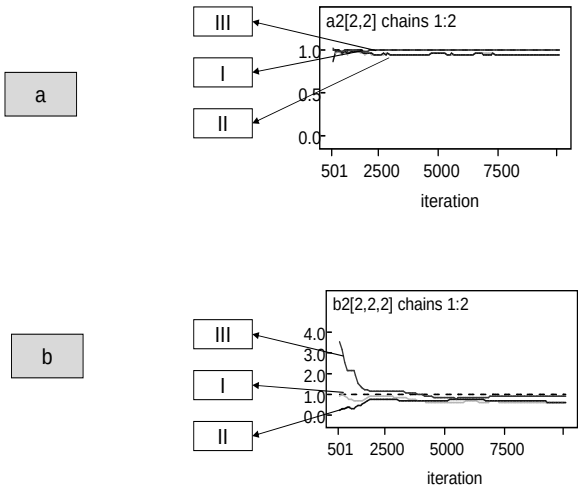
The model without covariates was chosen to illustrate the convergence analysis mentioned above. Figure 4.4 represents the plot of Gelman-Rubin statistic for the standard deviation of ability (i.e., a2[2,2] in Figure 4.4) for a class and item difficulty (i.e., b2[2,2,2] in Figure 4.4) for a item as an example. Patterns from the other estimates were similar to that of the estimates in Figure 4.4.

This method compared the within and between chain variances for each variable. In the plot, line I presents the width of the central 80% interval of the pooled runs. Line II represents the average width of the 80% interval within the individual runs. Line III represents the ratio of pooled over within (= R). When the chains have converged, the variance within each sequence and the variance between sequences for each variable will be roughly equal, so R should approximately equal one. Since R is almost equal to 1 from 5000th iteration for all other parameters shown in Figure 4.4), the chains were converged.

Geweke's test is based on a time-series analysis approach. It splits the sample into two parts, say after the first 10% and the last 50%. If the chain is at stationarity, the means of the two samples should be equal. A modified z-test, referred to as a Geweke z-score, is evaluated with a value larger than 2 indicating that the mean of the series is still drifting. When this is the case, a longer burn-in is required before the chain can be assumed to have reached stationarity. If the p-value for the Geweke z-score is less than .05 for all parameter estimates, the chains are assumed to be at stationarity.

The method of Raftery and Lewis is based on how many iterations are necessary to estimate the posterior for a given quantity. Here, a particular quantile (q) of the distribution of interest (typically 2.5% and 97.5%, yielding a 95% confidence interval), accuracy

Figure 4.4: Plot of Gelman-Rubin Statistic: (a) For Standard Deviation of Ability of a Class, (b) For Item Difficulty for a Item



of the quantile, and power for achieving this accuracy on the specified quantile need to be specified. With these three parameters set, the Raftery-Lewis test breaks the chain into a $(1, 0)$ sequence. This generates a two-state Markov chain, and the Raftery-Lewis test uses the sequence to estimate the transition probabilities. With these probabilities in hand, the number of addition burn-ins required to approach stationarity and the total chain length required to achieve the preset level of accuracy can be estimated. Results for the Raftery-Lewis test indicated less than 1,000 iterations were needed for the burn-in for all parameter estimates.

Based on results from these different methods, a conservative burn-in for all parameters was set at 7,000 iterations. An additional 8,000 iterations were run after discarding the burn-in iterations in order to obtain the posterior estimates.

Fitting MMixIRTM: Empirical Results

Class Proportions. Table 4.12 presents student-level class proportions for each school-level class, k. There was a similar pattern in proportions with the use of covariates and without. Without covariates, school-level class 1 had student-level class 2 and class 4 as dominant groups while school-level class 2 had student-level class 1 as a dominant group. Using covariates, school-level class 1 had student-level class 4 as a dominant group, while school-level class 2 had student-level class 1 as a dominant group.

Ability Comparisons. Both with and without covariates, student-level latent classes 1, 2, 3, and 4 were "Average", "Low", "High", and "Very Low" ability groups, respectively, as shown in Table 4.13. It appears that school-level class 1 was a low ability group while school-level class 2 was a high ability group.

Item Difficulty Profile. Class-specific item difficulty estimates are shown in Table 4.14 for both with and without covariates models. Correlations between item difficulty estimates without covariates and item difficulty estimates with covariates was .973. Different item difficulty profiles provide information on how each item functioned for each class. Table 4.14 shows both student- and school-level item profiles. In the DIF analysis, we compared item difficulties in two ways: (1) student-level (i.e., g) item difficulty comparison within a school-level class (i.e., k) for student-level DIF analysis, and (2) school-level (i.e., k) item difficulty comparison given a student-level class (i.e., g) for school-level DIF analysis.

DIF Detection. As previously indicated, hypotheses can be tested using the HPD interval. In this study, the 95-percent HPD interval was used. Tables 4.15, 4.16, 4.17, 4.18, 4.19, and 4.20 show DIF items detected at both student- and school-levels. Values inside parenthesis in those tables indicates the lower bound of HPD and upper bound of HPD. Items with bolded entries were detected as DIF items.

Tables 4.15, 4.16, and 4.17 present the DIF items detected for the model without covariates. As mentioned earlier, school-level DIF item(s) can be detected by comparing student-level item difficulties across school-level latent classes. At the school-level, shown in Table

4.15, there was 1 item at student-level class 1, 6 items at student-level class 2, 2 items at student-level class 3, and 7 items at student-level class 4 that were detected as DIF items. In addition, student-level DIF item(s) can be detected by comparing pairwise item difficulties within each school-level class. At school-level class 1 (i.e., $K = 1$) shown in 4.16, the number of DIF items varied from 10 to 30 across pairwise comparison of difficulties at the student-level class. At school-level class 2 (i.e., $K = 2$) shown in Table 4.17, the number of DIF items varied from 16 to 30 across pairwise comparison of difficulties at the student-level class.

Tables 4.18, 4.19, and 4.20 show the DIF items detected for the model with covariates. At the school-level shown in Table 4.18, there were 13 items at student-level class 1, 7 items at student-level class 2, 13 items at student-level class 3, and 2 items at student-level class 4 detected as DIF items. At school-level class 1 (i.e., $K = 1$) shown in Table 4.19, the number of DIF items varied from 12 to 29 across pairwise comparison of difficulties at the student-level class. At school-level class 2 (i.e., $K = 2$) shown in Table 4.20, the number of DIF items varied from 13 to 33 across pairwise comparison of difficulties at the student-level class.

Since looking at differential item functioning using a latent class approach should maximize the differences between groups, the large differences in item difficulties are expected (Samuelsen, 2005). The number of DIF items mentioned above based on a latent class group is larger than the what one would expect based on DIF analysis using manifest groups. This result is consistent with that previous research based on a latent class for DIF analysis (Cohen & Bolt, 2005; Cohen, Gregg, & Deng, 2005; Samuelsen, 2005).

DIF information shown in Tables 4.15, 4.16, 4.17, 4.18, 4.19, and 4.20 provides information about the characteristics of latent classes with respect to item characteristics at both student- and school-levels. One approach is to have individuals with content expertise examine the items and categorize them based on differences such as the kinds of cognitive skills required or perhaps surface level characteristics of the items involved (see, for example, Li, Cohen, & Ibarra, 2004). The categorizations of cognitive skills required by the items are shown in Table

4.21 as a Q-matrix (Tatsuoka, 1983). This Q-matrix is constructed based on skill information provided from the College Board for the PSAT/NMSQT. In Table 4.21, 1 was assigned if the skill was required by an item and 0 otherwise.

Table 4.9: Generating Item Difficulty Parameters and Their Transformed Estimates: Condition of 30% DIF and 25 Students/320 Schools without Covariate Model

	$G=1, K=1$		$G=2, K=1$		$G=1, K=2$		$G=2, K=2$	
Item	Parameter	Estimates	Parameter	Estmates	Parameter	Estimates	Parameter	Estimates
1	-2.00	-2.07	0.00	0.16	-2.40	-2.14	0.40	0.38
2	-2.00	-2.04	0.00	-0.12	-3.20	-3.40	1.20	1.26
3	-1.75	-1.75	0.00	0.10	-2.55	-2.70	0.80	0.73
4	-1.75	-1.68	0.25	0.30	-2.75	-2.87	1.25	1.24
5	-1.50	-1.45	0.25	0.11	-2.10	-1.95	0.85	0.86
6	-1.50	-1.48	0.25	0.32	-2.70	-2.71	1.45	1.45
7	-1.25	-1.21	0.25	0.28	-1.65	-1.68	0.65	0.63
8	-1.25	-1.14	0.50	0.44	-2.45	-2.18	1.70	1.63
9	-1.00	-0.98	0.50	0.73	-1.80	-1.69	1.30	1.31
10	-1.00	-0.96	0.50	0.61	-2.00	-2.05	1.50	1.50
11	-0.50	-0.51	1.00	0.93	-1.10	-1.10	1.60	1.63
12	-0.50	-0.45	1.00	1.18	-1.70	-1.77	2.20	2.17
13	-0.50	-0.50	1.00	0.95	-0.50	-0.49	1.00	1.00
14	-0.25	-0.26	1.00	0.82	-0.25	-0.31	1.00	1.00
15	-0.25	-0.29	1.25	1.22	-0.25	-0.25	1.25	1.21
16	-0.25	-0.31	1.25	1.33	-0.25	-0.30	1.25	1.23
17	-0.25	-0.25	1.25	1.33	-0.25	-0.16	1.25	1.26
18	0.50	0.55	1.25	1.14	0.50	0.42	1.25	1.30
19	0.50	0.44	1.50	1.29	0.50	0.53	1.50	1.55
20	0.50	0.49	1.50	1.43	0.50	0.43	1.50	1.54
21	0.00	-0.01	-0.50	-0.53	0.00	-0.11	-0.50	-0.63
22	0.00	-0.06	-0.50	-0.58	0.00	-0.05	-0.50	-0.46
23	0.00	-0.04	-0.50	-0.54	0.00	0.04	-0.50	-0.51
24	0.25	0.26	-0.25	-0.35	0.25	0.28	-0.25	-0.28
25	0.25	0.33	-0.25	-0.27	0.25	0.33	-0.25	-0.30
26	0.25	0.16	-0.25	-0.24	0.25	0.31	-0.25	-0.19
27	0.25	0.21	-0.25	-0.28	0.25	0.29	-0.25	-0.29
28	0.50	0.51	0.50	0.75	0.50	0.57	0.50	0.48
29	0.50	0.46	0.50	0.56	0.50	0.51	0.50	0.51
30	0.50	0.49	0.50	0.60	0.50	0.49	0.50	0.40
31	1.00	0.98	-2.00	-1.89	1.00	0.96	-2.00	-2.08
32	1.00	1.07	-2.00	-1.95	1.00	0.97	-2.00	-2.02
33	1.00	1.10	-1.75	-1.76	1.00	1.04	-1.75	-1.70
34	1.00	0.95	-1.75	-1.69	1.00	0.95	-1.75	-1.80
35	1.25	1.24	-1.50	-1.19	1.25	1.29	-1.50	-1.38
36	1.25	1.20	-1.50	-1.72	1.25	1.26	-1.50	-1.47
37	1.25	1.29	-1.25	-1.34	1.25	1.27	-1.25	-1.35
38	1.25	1.20	-1.25	-1.12	1.25	1.14	-1.25	-1.21
39	1.50	1.50	-1.00	-1.19	1.50	1.44	-1.00	-1.10
40	1.50	1.52	-1.00	-1.34	1.50	1.49	-1.00	-0.60

Table 4.10: Model Selection Result for Mathematics Section: Without Covariate Model

Number of Mixtures	Model	Npar	BIC	AIC
$G = 1$ $K = 1$	Multilevel Rasch Model	40	310200(0)	309900(0)
$G = 2$ $K = 2$	MMixIRTM	162	300500(0)	299300(0)
$G = 3$ $K = 2$	MMixIRTM	244	299000(0.12)	297300(0)
$G = 4$ $K = 2$	MMixIRTM	326	**298800(0.88)**	296500(0.25)
$G = 5$ $K = 2$	MMixIRTM	408	299500(0)	**296400(0.67)**
$G = 3$ $K = 3$	MMixIRTM	367	300200(0)	297200(0.02)
$G = 4$ $K = 3$	MMixIRTM	490	300800(0)	297300(0.01)
$G = 5$ $K = 3$	MMixIRTM	613	301200(0)	296900(0.05)

Table 4.11: Model Selection Result for Mathematics Section: With Covariate Model

Number of Mixtures	Model	Npar	BIC	AIC
$G = 1$ $K = 1$	Multilevel Rasch Model	40	310200(0)	309900(0)
$G = 2$ $K = 2$	MMixIRTM	182	300800(0)	299500(0)
$G = 3$ $K = 2$	MMixIRTM	266	299700(0.05)	297800(0)
$G = 4$ $K = 2$	MMixIRTM	370	**299400(0.64)**	296800(0.03)
$G = 5$ $K = 2$	MMixIRTM	438	299500(0.31)	**296400(0.93)**
$G = 3$ $K = 3$	MMixIRTM	409	301900(0)	299100(0)
$G = 4$ $K = 3$	MMixIRTM	545	301400(0)	297600(0)
$G = 5$ $K = 3$	MMixIRTM	681	301700(0)	296900(0.04)

Table 4.12: Class Proportions for Mathematics Section Within School-Level Group Membership

		$\hat{\pi}_k$	$G = 1$	$G = 2$	$G = 3$	$G = 4$
Without	$K = 1$	.462	.187	**.386**	.058	**.369**
Covariates	$K = 2$	.538	**.404**	.272	.194	.130
With	$K = 1$	.465	.246	.272	.130	**.351**
Covariates	$K = 2$	.535	**.396**	.279	.100	.224

Table 4.13: Distribution of Ability for Mathematics Section: With and Without Covariates

		$G=1$	$G=2$	$G=3$	$G=4$
Without	$K=1$	$N(0*, 0.572^2)$	$N(-1.965, 0.431^2)$	$N(1.344, 0.893^2)$	$N(-4.557, 0.299^2)$
Covariate	$K=2$	$N(0.562, 0.575^2)$	$N(-1.768, 0.412^2)$	$N(1.474, 1.067^2)$	$N(-4.001, 0.319^2)$
With	$K=1$	$N(0*, 1.131^2)$	$N(-1.842, 0.425^2)$	$N(0.535, 0.564^2)$	$N(-4.524, 0.311^2)$
Covariate	$K=2$	$N(0.570, 0.563^2)$	$N(-1.730, 0.442^2)$	$N(1.621, 1.032^2)$	$N(-4.250, 0.307^2)$

* Fixed for identification

Table 4.14: Class-Specific Item Difficulty for for Mathematics Section

	Without Covariates								With Covariates							
	$K=1$				$K=2$				$K=1$				$K=2$			
Item	$G=1$	$G=2$	$G=3$	$G=4$	$G=1$	$G=2$	$G=3$	$G=4$	$G=1$	$G=2$	$G=3$	$G=4$	$G=1$	$G=2$	$G=3$	$G=4$
1	-2.745	-2.646	-1.610	-2.654	-2.345	-2.511	-2.320	-2.553	-2.521	-2.578	-2.132	-2.686	-1.967	-2.585	-2.411	-2.562
2	-2.072	-1.924	-1.906	-2.584	-2.105	-1.911	-2.070	-2.398	-2.273	-2.039	-1.981	-2.475	-2.103	-1.796	-2.015	-2.528
3	-2.415	-1.140	-2.063	-0.446	-2.650	-1.281	-3.167	-0.368	-2.680	-1.182	-2.698	-0.261	-2.558	-1.356	-2.858	-0.674
4	-1.376	-0.904	-1.057	-0.666	-1.358	-0.983	-0.681	-0.348	-1.535	-0.990	-0.855	-0.612	-0.697	-1.010	-1.274	-0.489
5	-2.003	-1.986	-1.486	-1.723	-1.861	-1.890	-1.423	-1.772	-1.967	-2.060	-1.375	-1.782	-1.397	-1.880	-1.912	-1.672
6	-1.595	-1.224	-1.069	-0.538	-1.583	-0.928	-1.679	-0.449	-1.688	-1.031	-1.485	-0.589	-1.550	-1.305	-1.487	-0.387
7	-1.357	-1.889	-0.859	-1.008	-1.081	-1.790	-0.815	-0.974	-0.940	-1.792	-1.007	-1.102	-0.844	-1.864	-1.151	-1.009
8	-1.110	-0.985	-0.950	-0.908	-1.294	-0.789	-1.065	-0.769	-1.011	-0.888	-1.183	-0.869	-0.857	-0.983	-1.440	-0.805
9	-0.892	-0.502	-1.093	-0.290	-1.193	-0.622	-1.170	-0.146	-0.157	-0.594	-1.218	-1.118	-0.349	-0.562	-1.222	-0.967
10	-0.529	-0.665	-0.923	-0.583	-0.661	-0.702	-0.287	-0.591	-0.619	-0.773	-0.352	-0.628	-0.560	-0.576	-0.649	-0.511
11	-0.052	-0.303	-0.916	-0.827	0.062	-0.266	-0.597	-0.589	0.202	-0.316	-0.431	-0.765	-0.786	-0.253	-0.073	-0.671
12	-0.466	-0.199	-0.389	-0.492	-0.384	-0.076	-0.244	-0.207	-0.431	-0.145	-0.395	-0.383	-0.115	-0.217	-0.392	-0.375
13	0.109	0.254	-0.539	-0.503	-0.062	-0.080	-0.400	-0.333	-0.318	-0.011	-0.359	-0.476	-0.548	0.227	0.040	-0.312
14	0.338	-0.086	0.396	-0.133	0.299	0.185	0.495	-0.185	0.264	0.039	0.348	-0.192	0.508	0.091	0.386	-0.098
15	0.298	0.942	-0.080	-0.636	0.194	0.832	0.230	-0.277	0.017	0.961	0.264	-0.449	-0.024	0.784	0.294	-0.481
16	2.101	1.248	1.672	-0.198	2.442	1.215	1.880	-0.102	2.464	1.145	1.825	-0.043	1.762	1.536	2.334	-0.198
17	1.561	0.736	2.169	-0.137	1.460	0.737	1.920	-0.178	1.304	0.698	1.944	-0.005	1.873	0.938	1.672	-0.242
18	2.454	0.712	2.601	-0.267	2.032	0.737	2.434	-0.071	2.270	0.745	2.430	-0.127	2.333	0.887	2.184	-0.227
19	2.425	1.389	2.400	-0.310	2.403	1.560	2.339	0.184	2.259	1.716	2.324	-0.146	2.270	1.484	2.571	-0.063
20	2.172	1.459	1.970	0.280	2.085	1.488	1.909	0.388	2.138	1.503	1.887	0.459	1.892	1.655	2.073	0.226
21	-1.721	-1.233	-1.201	-0.139	-1.467	-1.302	-1.081	-0.034	-1.476	-1.209	-1.119	-0.256	-1.177	-1.489	-1.508	0.017
22	-2.640	-1.907	-2.084	-0.730	-2.166	-1.880	-1.900	-0.740	-2.444	-1.886	-1.868	-0.782	-1.887	-2.076	-2.170	-0.777
23	-0.763	-0.022	-0.643	0.112	-0.778	0.083	-0.395	0.144	-0.692	-1.110	-0.528	0.316	-0.503	-0.079	-0.784	-0.024
24	-1.180	-0.652	-0.839	-0.141	-1.261	-0.678	-1.754	-0.170	-1.093	-0.707	-1.661	-0.153	-1.177	-0.721	-1.392	-0.166
25	0.037	0.082	-0.479	0.228	-0.151	0.084	-0.150	0.695	-0.046	0.261	-0.204	0.287	-0.303	-0.084	-0.143	0.385
26	1.506	0.603	2.003	-0.321	1.440	0.768	2.252	0.068	1.427	0.678	2.075	-0.216	2.233	0.843	1.537	-0.142
27	1.526	0.975	0.613	0.243	1.529	0.957	1.004	-0.055	1.519	1.187	0.992	0.166	0.854	0.911	1.478	0.213
28	1.535	0.446	1.396	-0.510	1.775	0.736	1.350	-0.743	1.869	0.692	1.285	-0.514	1.358	0.655	1.686	-0.591
29	0.296	-0.042	-1.123	-0.807	0.297	-0.329	-0.721	-0.532	0.456	-0.120	-0.575	-0.685	-0.868	-0.153	0.178	-0.706
30	-0.387	0.107	0.131	1.556	-0.155	0.030	0.189	1.149	-0.211	-0.048	-0.003	1.170	0.348	0.063	-0.231	1.633
31	-1.157	-1.125	-0.740	0.284	-1.148	-0.901	-1.124	0.271	-1.190	-1.060	-1.032	0.358	-0.889	-1.192	-1.283	0.172
32	-0.869	-0.041	-0.989	1.464	-1.047	-0.229	-0.989	1.141	-1.200	-0.172	-0.787	1.247	-1.175	-0.297	-0.919	1.360
33	-0.496	-0.161	-0.245	1.246	-0.463	-0.144	-0.293	1.321	-0.478	-0.234	-0.386	1.270	-0.339	-0.213	-0.404	1.181
34	1.158	0.533	1.896	-0.086	1.374	0.453	2.045	-0.163	1.299	0.528	1.949	0.007	1.917	0.589	1.427	-0.222
35	3.101	3.601	1.385	3.671	2.725	3.572	1.181	3.263	3.059	3.880	1.343	3.577	1.017	3.482	2.632	3.474
36	1.321	1.047	1.294	2.997	1.316	1.166	1.569	2.136	1.375	1.078	1.406	2.533	1.527	0.998	1.394	2.651
37	1.599	1.187	1.935	1.534	1.569	1.359	2.225	0.766	1.545	1.055	2.128	1.240	2.099	1.509	1.655	1.244
38	2.699	4.313	1.421	4.024	2.212	3.330	1.305	3.221	2.292	3.778	1.432	3.712	1.251	4.036	2.176	3.772

Table 4.15: Magnitude of School-Level DIF Values for Mathematics Section: Without Covariates

	Across School-Level Latent Classes			
Item	Student-Level Class 1	Student-Level Class 2	Student-Level Class 3	Student-Level Class4
1	-0.400 (-0.944, 0.096)	-0.135 (-0.409, 0.140)	0.710 (-0.160, 1.617)	-0.101 (-0.3644, 0.1543)
2	0.033 (-0.401, 0.456)	-0.012 (-0.235, 0.217)	0.165 (-0.791, 1.124)	-0.186 (-0.448 , 0.08163)
3	0.234 (-0.430, 0.849)	0.141 (-0.100, 0.384)	1.104 (-0.041, 2.304)	-0.077 (-0.4136, 0.2386)
4	-0.018 (-0.364, 0.316)	0.079 (-0.134, 0.296)	-0.376 (-1.067, 0.324)	**-0.318 (-0.6132, -0.0276)**
5	-0.142 (-0.564, 0.251)	-0.096 (-0.328, 0.134)	-0.063 (-0.845, 0.685)	0.049 (-0.1984, 0.2948)
6	-0.012 (-0.366, 0.334)	**-0.296 (-0.514, -0.071)**	0.610 (-0.138, 1.361)	-0.089 (-0.3819, 0.1951)
7	-0.276 (-0.607, 0.042)	-0.099 (-0.340, 0.140)	-0.045 (-0.694, 0.588)	-0.034 (-0.3402, 0.2432)
8	0.184 (-0.131, 0.514)	-0.196 (-0.405, 0.021)	0.115 (-0.576, 0.780)	-0.139 (-0.4146, 0.1394)
9	0.301 (-0.007, 0.612)	0.120 (-0.105, 0.341)	0.077 (-0.715, 0.769)	-0.144 (-0.4627, 0.1608)
10	0.133 (-0.131, 0.395)	0.037 (-0.173, 0.242)	**-0.636 (-1.345, -0.040)**	0.008 (-0.2778, 0.2994)
11	-0.114 (-0.380, 0.149)	-0.038 (-0.254, 0.183)	-0.319 (-1.202, 0.411)	-0.238 (-0.5167, 0.0329)
12	-0.082 (-0.342, 0.176)	-0.123 (-0.356, 0.110)	-0.145 (-0.706, 0.387)	-0.285 (-0.5922, 0.01118)
13	0.171 (-0.092, 0.434)	**0.334 (0.096, 0.578)**	-0.138 (-0.940, 0.451)	-0.170 (-0.4547, 0.1145)
14	0.039 (-0.220, 0.312)	**-0.270 (-0.505, -0.041)**	-0.099 (-0.579, 0.355)	0.051 (-0.2555, 0.3563)
15	0.104 (-0.186, 0.386)	0.110 (-0.208, 0.416)	-0.310 (-0.833, 0.190)	**-0.359 (-0.6692, -0.0523)**
16	-0.342 (-0.711, 0.046)	0.033 (-0.330, 0.395)	-0.208 (-0.655, 0.218)	-0.096 (-0.4, 0.2039)
17	0.101 (-0.241, 0.437)	-0.001 (-0.299, 0.287)	0.249 (-0.163, 0.677)	0.041 (-0.2594, 0.339)
18	0.422 (-0.127, 0.981)	-0.025 (-0.322, 0.261)	0.167 (-0.274, 0.597)	-0.196 (-0.5175, 0.1112)
19	0.022 (-0.408, 0.468)	-0.171 (-0.543, 0.198)	0.061 (-0.379, 0.476)	**-0.494 (-0.8123, -0.1651)**
20	0.087 (-0.277, 0.460)	-0.029 (-0.405, 0.334)	0.062 (-0.393, 0.478)	-0.108 (-0.458 , 0.2501)
21	-0.254 (-0.635, 0.105)	0.069 (-0.168, 0.309)	-0.120 (-0.833, 0.564)	-0.105 (-0.4793, 0.2502)
22	-0.474 (-1.051, 0.052)	-0.026 (-0.289, 0.239)	-0.184 (-1.262, 0.836)	0.009 (-0.2963, 0.3037)
23	0.015 (-0.267, 0.299)	-0.105 (-0.363, 0.173)	-0.248 (-0.823, 0.326)	-0.032 (-0.3722, 0.2982)
24	0.081 (-0.243, 0.393)	0.027 (-0.194, 0.251)	**0.915 (0.182, 1.652)**	0.029 (-0.305, 0.3517)
25	0.188 (-0.079, 0.457)	-0.001 (-0.236, 0.228)	-0.329 (-1.108, 0.263)	**-0.467 (-0.9077, -0.062)**
26	0.066 (-0.257, 0.390)	-0.165 (-0.466, 0.116)	-0.249 (-0.662, 0.179)	**-0.389 (-0.7019, -0.0794)**
27	-0.003 (-0.344, 0.401)	0.018 (-0.300, 0.331)	-0.392 (-0.883, 0.060)	0.298 (-0.0201, 0.6174)
28	-0.239 (-0.565, 0.110)	-0.290 (-0.609, 0.027)	0.046 (-0.366, 0.465)	0.233 (-0.0407, 0.5137)
29	-0.001 (-0.299, 0.310)	**0.287 (0.048, 0.518)**	-0.402 (-1.298, 0.378)	**-0.275 (-0.5471, -0.0118)**
30	-0.232 (-0.516, 0.044)	0.077 (-0.170, 0.331)	-0.059 (-0.564, 0.461)	0.407 (-0.2039, 0.9806)
31	**-0.419 (-0.764, -0.096)**	**-0.224 (-0.445, -0.012)**	0.384 (-0.274, 1.072)	0.013 (-0.433, 0.4189)
32	0.178 (-0.139, 0.493)	0.188 (-0.069, 0.447)	0.000 (-0.710, 0.697)	0.323 (-0.2898, 0.8889)
33	-0.033 (-0.335, 0.251)	-0.018 (-0.253, 0.216)	0.049 (-0.581, 0.586)	-0.075 (-0.7754, 0.5585)
34	-0.216 (-0.576, 0.132)	0.079 (-0.185, 0.343)	-0.149 (-0.557, 0.253)	0.077 (-0.2354, 0.3758)
35	0.376 (-0.218, 1.039)	0.029 (-0.880, 0.928)	0.204 (-0.368, 0.707)	0.408 (-0.7822, 1.584)
36	0.006 (-0.300, 0.309)	-0.119 (-0.436, 0.183)	-0.275 (-0.704, 0.100)	0.861 (-0.0611, 1.795)
37	0.030 (-0.321, 0.396)	-0.172 (-0.507, 0.166)	-0.290 (-0.715, 0.109)	**0.768 (0.2857 , 1.269)**
38	0.487 (-0.006, 1.106)	**0.983 (0.031, 1.964)**	0.116 (-0.416, 0.562)	0.803 (-0.4337, 2.025)
Number of DIF Items	1	6	2	7

Table 4.16: Magnitude of Student-Level DIF Values for Mathematics Section: Without Covariates Model for $K = 1$

Item	$G = 1 Vs. G = 2$	$G = 1 Vs. G = 3$	$G = 1 Vs. G = 4$	$G = 2 Vs. G = 3$	$G = 2 Vs. G = 4$	$G = 3 Vs. G = 4$
1	-0.099 (-0.666, 0.387)	**-1.135 (-1.991, -0.233)**	-0.091 (-0.586, 0.349)	**-1.036 (-1.730, -0.334)**	0.008 (-0.233, 0.249)	**1.044 (0.342, 1.710)**
2	-0.149 (-0.559, 0.235)	-0.167 (-1.012, 0.764)	**0.512 (0.142, 0.847)**	-0.018 (-0.750, 0.799)	**0.661 (0.439, 0.885)**	0.679 (-0.138, 1.397)
3	**-1.275 (-1.869, -0.779)**	-0.353 (-1.433, 0.711)	**-1.970 (-2.521, -1.511)**	**0.923 (0.121, 1.812)**	**-0.694 (-0.938, -0.455)**	**-1.617 (-2.511, -0.816)**
4	**-0.472 (-0.822, -0.136)**	-0.319 (-1.108, 0.440)	**-0.710 (-1.027, -0.411)**	0.153 (-0.501, 0.787)	**-0.238 (-0.463, -0.015)**	-0.391 (-1.026, 0.265)
5	-0.017 (-0.438, 0.367)	-0.518 (-1.276, 0.281)	-0.280 (-0.642, 0.049)	-0.501 (-1.154, 0.178)	**-0.263 (-0.483, -0.044)**	0.238 (-0.440, 0.885)
6	**-0.371 (-0.739, -0.022)**	-0.526 (-1.249, 0.214)	**-1.057 (-1.383, -0.743)**	-0.155 (-0.763, 0.486)	**-0.686 (-0.922, -0.461)**	-0.532 (-1.165, 0.087)
7	**0.531 (0.188, 0.867)**	-0.498 (-1.111, 0.177)	**-0.349 (-0.659, -0.046)**	**-1.030 (-1.560, -0.455)**	**-0.881 (-1.123, -0.645)**	0.149 (-0.417, 0.692)
8	-0.126 (-0.444, 0.178)	-0.160 (-0.831, 0.540)	-0.202 (-0.479, 0.071)	-0.035 (-0.603, 0.561)	-0.076 (-0.292, 0.139)	-0.041 (-0.639, 0.537)
9	**-0.390 (-0.709, -0.076)**	0.201 (-0.462, 0.984)	**-0.602 (-0.890, -0.326)**	**0.591 (0.018, 1.300)**	-0.212 (-0.453, 0.031)	**-0.803 (-1.526, -0.223)**
10	0.136 (-0.160, 0.418)	0.394 (-0.221, 1.118)	0.054 (-0.206, 0.300)	0.258 (-0.281, 0.946)	-0.082 (-0.313, 0.146)	-0.340 (-1.022, 0.210)
11	0.251 (-0.040, 0.541)	**0.863 (0.180, 1.761)**	**0.775 (0.522, 1.029)**	**0.612 (0.007, 1.453)**	**0.524 (0.307, 0.738)**	-0.088 (-0.943, 0.522)
12	-0.267 (-0.571, 0.021)	-0.077 (-0.616, 0.526)	0.026 (-0.233, 0.282)	0.190 (-0.282, 0.695)	**0.293 (0.066, 0.534)**	0.103 (-0.429, 0.575)
13	-0.145 (-0.449, 0.160)	**0.648 (0.059, 1.437)**	**0.612 (0.342, 0.881)**	**0.793 (0.275, 1.561)**	**0.757 (0.524, 0.992)**	-0.036 (-0.814, 0.485)
14	**0.423 (0.137, 0.708)**	-0.058 (-0.540, 0.470)	**0.471 (0.199, 0.746)**	**-0.481 (-0.888, -0.033)**	0.048 (-0.214, 0.287)	**0.529 (0.077, 0.955)**
15	**-0.644 (-1.003, -0.280)**	0.378 (-0.202, 0.960)	**0.934 (0.625, 1.250)**	**1.022 (0.538, 1.522)**	**1.578 (1.303, 1.865)**	**0.556 (0.070, 1.002)**
16	**0.853 (0.431, 1.290)**	0.429 (-0.093, 0.991)	**2.298 (1.972, 2.656)**	-0.424 (-0.877, 0.079)	**1.446 (1.152, 1.737)**	**1.869 (1.422, 2.276)**
17	**0.825 (0.461, 1.173)**	**-0.608 (-1.081, -0.100)**	**1.698 (1.383, 2.030)**	**-1.433 (-1.840, -1.025)**	**0.873 (0.611, 1.131)**	**2.306 (1.913, 2.711)**
18	**1.742 (1.224, 2.277)**	-0.148 (-0.834, 0.624)	**2.721 (2.231, 3.256)**	**-1.889 (-2.312, -1.435)**	**0.979 (0.711, 1.234)**	**2.868 (2.435, 3.286)**
19	**1.036 (0.577, 1.513)**	0.025 (-0.553, 0.675)	**2.735 (2.349, 3.147)**	**-1.011 (-1.443, -0.556)**	**1.699 (1.402, 1.989)**	**2.710 (2.281, 3.113)**
20	**0.713 (0.282, 1.136)**	0.202 (-0.352, 0.799)	**1.892 (1.534, 2.255)**	**-0.511 (-0.956, -0.018)**	**1.179 (0.862, 1.501)**	**1.690 (1.229, 2.101)**
21	**-0.488 (-0.878, -0.126)**	-0.520 (-1.259, 0.234)	**-1.582 (-1.933, -1.251)**	-0.032 (-0.652, 0.610)	**-1.094 (-1.359, -0.836)**	**-1.062 (-1.712, -0.426)**
22	**-0.733 (-1.278, -0.243)**	-0.556 (-1.646, 0.571)	**-1.909 (-2.416, -1.472)**	0.177 (-0.676, 1.100)	**-1.177 (-1.422, -0.931)**	**-1.353 (-2.273, -0.493)**
23	**-0.741 (-1.078, -0.427)**	-0.120 (-0.742, 0.497)	**-0.874 (-1.178, -0.581)**	**0.621 (0.088, 1.143)**	-0.133 (-0.409, 0.159)	**-0.754 (-1.296, -0.215)**
24	**-0.529 (-0.862, -0.210)**	-0.341 (-0.977, 0.331)	**-1.039 (-1.338, -0.755)**	**0.187 (0.378, 0.778)**	**-0.510 (-0.753, -0.272)**	**-0.698 (-1.276, -0.146)**
25	-0.045 (-0.345, 0.257)	0.516 (-0.114, 1.329)	-0.191 (-0.474, 0.087)	**0.561 (0.020, 1.324)**	-0.145 (-0.433, 0.129)	**-0.706 (-1.484, -0.135)**
26	**0.903 (0.552, 1.238)**	-0.497 (-0.984, 0.011)	**1.827 (1.515, 2.138)**	**-1.400 (-1.801, -1.001)**	**0.924 (0.651, 1.182)**	**2.324 (1.937, 2.733)**
27	**0.551 (0.131, 1.026)**	**0.914 (0.371 , 1.519)**	**1.283 (0.928, 1.689)**	0.363 (-0.077, 0.859)	**0.732 (0.440, 1.028)**	0.369 (-0.117, 0.807)
28	**1.089 (0.733, 1.466)**	0.140 (-0.372, 0.686)	**2.045 (1.749, 2.369)**	**-0.950 (-1.350, -0.530)**	**0.956 (0.695, 1.213)**	**1.906 (1.503, 2.305)**
29	0.338 (-0.001, 0.699)	**1.419 (0.728 , 2.229)**	**1.103 (0.811, 1.404)**	**1.081 (0.435, 1.888)**	**0.765 (0.538, 0.994)**	-0.316 (-1.110, 0.311)
30	**-0.494 (-0.823, -0.178)**	-0.517 (-1.066, 0.036)	**-1.943 (-2.386, -1.543)**	-0.024 (-0.528, 0.449)	**-1.449 (-1.878, -1.053)**	**-1.425 (-2.019, -0.828)**
31	**-0.442 (-0.812, -0.102)**	**-0.827 (-1.512, -0.172)**	**-1.851 (-2.227, -1.508)**	-0.385 (-1.019, 0.188)	**-1.409 (-1.689, -1.135)**	**-1.024 (-1.621, -0.377)**
32	**-0.829 (-1.157, -0.509)**	0.120 (-0.601, 0.850)	**-2.333 (-2.767, -1.922)**	**0.948 (0.317, 1.593)**	**-1.505 (-1.926, -1.124)**	**-2.453 (-3.172, -1.765)**
33	**-0.334 (-0.672, -0.014)**	-0.251 (-0.850, 0.463)	**-1.742 (-2.195, -1.346)**	0.083 (-0.402, 0.660)	**-1.407 (-1.831, -1.029)**	**-1.491 (-2.144, -0.914)**
34	**0.626 (0.228, 0.992)**	**-0.738 (-1.232, -0.214)**	**1.244 (0.900, 1.584)**	**-1.364 (-1.749, -0.982)**	**0.619 (0.351, 0.886)**	**1.982 (1.586, 2.365)**
35	-0.500 (-1.420, 0.401)	**1.716 (1.081, 2.415)**	-0.570 (-1.549, 0.349)	**2.216 (1.531, 3.058)**	-0.069 (-1.109, 0.952)	**-2.285 (-3.274, -1.442)**
36	0.275 (-0.101, 0.631)	0.028 (-0.448, 0.547)	**-1.675 (-2.442, -1.016)**	-0.247 (-0.621, 0.153)	**-1.950 (-2.714, -1.300)**	**-1.703 (-2.513, -1.009)**
37	0.412 (-0.012, 0.819)	-0.337 (-0.861, 0.235)	0.065 (-0.391, 0.504)	**-0.748 (-1.147, -0.336)**	-0.347 (-0.803, 0.066)	0.401 (-0.116, 0.884)
38	**-1.614 (-2.525, -0.682)**	**1.278 (0.707, 1.899)**	**-1.325 (-2.353, -0.363)**	**2.892 (2.074, 3.813)**	0.289 (-0.857, 1.425)	**-2.603 (-3.639, -1.698)**
Number of DIF Items	24	10	30	23	28	25

Table 4.17: Magnitude of Student-Level DIF Values for Mathematics Section: Without Covariates for $K = 2$

Item	$G = 1 Vs. G = 2$	$G = 1 Vs. G = 3$	$G = 1 Vs. G = 4$	$G = 2 Vs. G = 3$	$G = 2 Vs. G = 4$	$G = 3 Vs. G = 4$
1	0.166 (-0.196, 0.518)	-0.025 (-0.680, 0.701)	0.208 (-0.113, 0.522)	-0.191 (-0.746, 0.456)	0.042 (-0.301, 0.370)	0.233 (-0.420, 0.815)
2	-0.194 (-0.528, 0.114)	-0.035 (-0.683, 0.704)	0.293 (-0.036, 0.613)	0.159 (-0.367, 0.798)	0.487 (0.190, 0.784)	0.328 (-0.310, 0.881)
3	**-1.368 (-1.769, -1.004)**	0.518 (-0.385, 1.553)	**-2.281 (-2.708, -1.891)**	**1.886 (1.121, 2.818)**	**-0.912 (-1.262, -0.581)**	**-2.799 (-3.749, -2.012)**
4	**-0.375 (-0.640, -0.113)**	**-0.677 (-1.066, -0.273)**	**-1.010 (-1.322, -0.709)**	-0.302 (-0.639, 0.036)	**-0.635 (-0.959, -0.325)**	-0.333 (-0.744, 0.046)
5	0.029 (-0.287, 0.334)	-0.439 (-0.939, 0.102)	-0.089 (-0.386, 0.199)	**-0.468 (-0.876, -0.023)**	-0.118 (-0.421, 0.180)	0.350 (-0.117, 0.777)
6	**-0.655 (-0.932, -0.389)**	0.096 (-0.421, 0.672)	**-1.134 (-1.431, -0.850)**	**0.751 (0.309, 1.253)**	**-0.479 (-0.800, -0.161)**	**-1.230 (-1.771, -0.749)**
7	**0.708 (0.456, 0.968)**	-0.267 (-0.652, 0.156)	-0.108 (-0.427, 0.185)	**-0.975 (-1.324, -0.596)**	**-0.815 (-1.163, -0.502)**	0.159 (-0.246, 0.532)
8	**-0.506 (-0.778, -0.239)**	-0.229 (-0.679, 0.241)	**-0.525 (-0.814, -0.253)**	0.276 (-0.098, 0.689)	-0.019 (-0.320, 0.285)	-0.296 (-0.727, 0.110)
9	**-0.572 (-0.834, -0.317)**	-0.024 (-0.491, 0.468)	**-1.047 (-1.365, -0.743)**	**0.548 (0.153, 0.965)**	**-0.475 (-0.820, -0.138)**	**-1.023 (-1.512, -0.568)**
10	0.040 (-0.189, 0.274)	**-0.375 (-0.708, -0.028)**	-0.071 (-0.351, 0.202)	**-0.415 (-0.716, -0.115)**	-0.111 (-0.416, 0.197)	0.304 (-0.053, 0.649)
11	**0.328 (0.088, 0.561)**	**0.659 (0.272, 1.111)**	**0.651 (0.388, 0.907)**	0.331 (-0.025, 0.748)	**0.323 (0.010, 0.630)**	-0.008 (-0.461, 0.405)
12	**-0.308 (-0.556, -0.071)**	-0.140 (-0.465, 0.201)	-0.177 (-0.456, 0.103)	0.168 (-0.132, 0.471)	**0.131 (-0.200, 0.462)**	-0.037 (-0.407, 0.326)
13	0.018 (-0.227, 0.257)	0.338 (-0.016, 0.698)	**0.271 (0.007, 0.535)**	0.320 (-0.008, 0.662)	**0.253 (-0.073, 0.575)**	-0.067 (-0.440, 0.303)
14	0.115 (-0.125, 0.341)	-0.196 (-0.476, 0.089)	**0.484 (0.207, 0.758)**	**-0.310 (-0.573, -0.043)**	**0.369 (0.048, 0.691)**	**0.679 (0.349, 1.003)**
15	**-0.638 (-0.922, -0.368)**	-0.036 (-0.336, 0.268)	**0.471 (0.182, 0.767)**	**0.602 (0.297, 0.916)**	**1.109 (0.758, 1.473)**	**0.506 (0.173, 0.847)**
16	**1.227 (0.893, 1.568)**	**0.563 (0.251, 0.888)**	**2.544 (2.227, 2.867)**	**-0.665 (-0.980, -0.338)**	**1.317 (0.950, 1.668)**	**1.982 (1.657, 2.296)**
17	**0.723 (0.444, 0.992)**	**-0.461 (-0.732, -0.177)**	**1.637 (1.342, 1.925)**	**-1.183 (-1.456, -0.912)**	**0.915 (0.571, 1.257)**	**2.098 (1.785, 2.410)**
18	**1.295 (0.998, 1.586)**	**-0.402 (-0.717, -0.073)**	**2.103 (1.777, 2.419)**	**-1.697 (-1.980, -1.411)**	**0.808 (0.437, 1.161)**	**2.505 (2.186, 2.824)**
19	**0.843 (0.489, 1.183)**	0.064 (-0.249, 0.387)	**2.219 (1.894, 2.544)**	**-0.779 (-1.110, -0.424)**	**1.376 (0.977, 1.791)**	**2.155 (1.814, 2.504)**
20	**0.597 (0.238, 0.935)**	0.176 (-0.121, 0.481)	**1.697 (1.365, 2.027)**	**-0.421 (-0.728, -0.094)**	**1.100 (0.693, 1.514)**	**1.521 (1.161, 1.879)**
21	-0.165 (-0.455, 0.112)	-0.386 (-0.835, 0.056)	**-1.432 (-1.799, -1.097)**	-0.221 (-0.605, 0.167)	**-1.267 (-1.637, -0.920)**	**-1.046 (-1.518, -0.594)**
22	-0.285 (-0.659, 0.076)	-0.265 (-0.935, 0.468)	**-1.426 (-1.795, -1.083)**	0.020 (-0.509, 0.610)	**-1.141 (-1.484, -0.813)**	**-1.161 (-1.804, -0.604)**
23	**-0.861 (-1.124, -0.602)**	**-0.382 (-0.735, -0.023)**	**-0.921 (-1.236, -0.621)**	**0.479 (0.144, 0.814)**	-0.060 (-0.424, 0.306)	**-0.539 (-0.924, -0.157)**
24	**-0.583 (-0.841, -0.325)**	0.493 (-0.054, 1.084)	**-1.091 (-1.413, -0.791)**	**1.076 (0.600, 1.608)**	**-0.507 (-0.859, -0.176)**	**-1.584 (-2.183, -1.044)**
25	-0.235 (-0.480, 0.003)	-0.001 (-0.333, 0.338)	**-0.846 (-1.242, -0.487)**	0.234 (-0.064, 0.540)	**-0.611 (-1.075, -0.195)**	**-0.845 (-1.314, -0.422)**
26	**0.672 (0.406, 0.932)**	**-0.812 (-1.073, -0.548)**	**1.372 (1.061, 1.671)**	**-1.484 (-1.751, -1.209)**	**0.700 (0.337, 1.054)**	**2.185 (1.855, 2.512)**
27	**0.572 (0.271, 0.868)**	**0.525 (0.228, 0.824)**	**1.584 (1.289, 1.886)**	-0.047 (-0.348, 0.264)	**1.012 (0.663, 1.364)**	**1.059 (0.710, 1.393)**
28	**1.039 (0.721, 1.356)**	**0.425 (0.139, 0.725)**	**2.518 (2.245, 2.807)**	**-0.614 (-0.899, -0.312)**	**1.479 (1.150, 1.810)**	**2.093 (1.785, 2.397)**
29	**0.626 (0.396, 0.854)**	**1.019 (0.598, 1.521)**	**0.829 (0.568, 1.080)**	0.393 (-0.016, 0.898)	0.203 (-0.106, 0.505)	-0.190 (-0.696, 0.235)
30	-0.186 (-0.441, 0.066)	**-0.344 (-0.650, -0.022)**	**-1.304 (-1.815, -0.863)**	-0.159 (-0.445, 0.136)	**-1.119 (-1.669, -0.637)**	**-0.960 (-1.498, -0.480)**
31	-0.246 (-0.504, 0.006)	-0.023 (-0.463, 0.437)	**-1.419 (-1.817, -1.065)**	0.223 (-0.158, 0.628)	**-1.172 (-1.571, -0.809)**	**-1.396 (-1.906, -0.919)**
32	**-0.819 (-1.104, -0.545)**	-0.058 (-0.489, 0.399)	**-2.188 (-2.708, -1.726)**	**0.761 (0.384 , 1.164)**	**-1.369 (-1.933, -0.863)**	**-2.130 (-2.764, -1.569)**
33	**-0.319 (-0.576, -0.079)**	-0.169 (-0.516, 0.195)	**-1.784 (-2.403, -1.279)**	0.150 (-0.155, 0.475)	**-1.465 (-2.110, -0.928)**	**-1.615 (-2.286, -1.038)**
34	**0.921 (0.677, 1.169)**	**-0.670 (-0.938, -0.398)**	**1.537 (1.242, 1.821)**	**-1.591 (-1.854, -1.332)**	**0.616 (0.274, 0.941)**	**2.207 (1.888, 2.520)**
35	**-0.847 (-1.635, -0.159)**	**1.544 (1.172, 1.938)**	-0.538 (-1.517, 0.317)	**2.391 (1.741, 3.138)**	**0.309 (0.778, 1.400)**	**-2.082 (-3.045, -1.236)**
36	0.150 (-0.162, 0.442)	-0.253 (-0.514, 0.011)	**-0.820 (-1.530, -0.230)**	**-0.403 (-0.680, -0.105)**	**-0.969 (-1.730, -0.316)**	-0.567 (-1.286, 0.037)
37	0.210 (-0.120, 0.524)	**-0.656 (-0.929, -0.387)**	**0.803 (0.418, 1.172)**	**-0.866 (-1.162, -0.563)**	**0.593 (0.119, 1.058)**	**1.459 (1.062, 1.851)**
38	**-1.118 (-1.789, -0.512)**	**0.907 (0.596, 1.223)**	**-1.009 (-1.967, -0.174)**	**2.025 (1.448, 2.699)**	0.109 (-1.007, 1.170)	**-1.916 (-2.862, -1.105)**
Number of DIF Items	25	16	31	23	30	26

Table 4.18: Magnitude of School-Level DIF Values for Mathematics Section: With Covariates

	Across School-Level Latent Classes			
Item	Student-Level Class 1	Student-Level Class 2	Student-Level Class 3	Student-Level Class 4
1	-0.5544(-1.293, 0.246)	0.007(-0.244, 0.264)	0.279(-0.337, 0.854)	-0.124(-0.342,0.099)
2	-0.1701(-0.933, 0.646)	**-0.242(-0.464, -0.028)**	0.034(-0.527, 0.558)	0.053(-0.210,0.303)
3	-0.1225(-1.054, 0.923)	0.175(-0.035, 0.388)	0.160(-0.671, 0.944)	**0.413(0.099,0.725)**
4	**-0.8382(-1.328, -0.343)**	0.019(-0.196, 0.234)	**0.420(0.023, 0.824)**	-0.123(-0.386,0.127)
5	-0.5704(-1.174, 0.049)	-0.181(-0.405, 0.048)	**0.537(0.061, 1.001)**	-0.110(-0.325,0.097)
6	-0.1373(-0.760, 0.548)	**0.274(0.077, 0.471)**	0.002(-0.468, 0.452)	-0.203(-0.472,0.061)
7	-0.0959(-0.582, 0.406)	0.072(-0.185, 0.314)	0.144(-0.258, 0.538)	-0.093(-0.325,0.134)
8	-0.1537(-0.645, 0.346)	0.094(-0.161, 0.330)	0.258(-0.211, 0.686)	-0.064(-0.298,0.171)
9	-0.1512(-0.652, 0.395)	-0.032(-0.264, 0.190)	0.005(-0.468, 0.433)	0.191(-0.065,0.462)
10	-0.1163(-0.542, 0.340)	**-0.197(-0.392, -0.008)**	0.298(-0.039, 0.639)	-0.059(-0.305,0.194)
11	**0.988(0.495, 1.572)**	-0.064(-0.262, 0.141)	**-0.359(-0.765, -0.013)**	-0.094(-0.337,0.148)
12	-0.3158(-0.704, 0.081)	0.073(-0.182, 0.316)	-0.003(-0.342, 0.336)	-0.007(-0.274,0.251)
13	0.4102(-0.024, 0.903)	-0.238(-0.481, 0.013)	**-0.399(-0.775, -0.065)**	-0.165(-0.422,0.104)
14	-0.2443(-0.586, 0.118)	-0.052(-0.273, 0.175)	-0.038(-0.323, 0.253)	-0.094(-0.361,0.167)
15	0.04067(-0.345, 0.450)	0.177(-0.129, 0.517)	-0.030(-0.319, 0.285)	0.032(-0.225,0.282)
16	**0.702(0.313, 1.118)**	**-0.391(-0.724, -0.061)**	**-0.510(-0.877, -0.199)**	0.154(-0.141,0.441)
17	**-0.5691(-0.904, -0.222)**	-0.240(-0.522, 0.054)	0.272(-0.034, 0.568)	0.236(-0.032,0.508)
18	-0.0631(-0.468, 0.353)	-0.142(-0.407, 0.130)	0.246(-0.095, 0.561)	0.100(-0.173,0.367)
19	-0.0114(-0.375, 0.357)	0.232(-0.145, 0.611)	-0.247(-0.603, 0.078)	-0.083(-0.383,0.218)
20	0.2455(-0.108, 0.628)	-0.152(-0.516, 0.206)	-0.187(-0.543, 0.114)	0.234(-0.093,0.568)
21	-0.299(-0.830, 0.272)	**0.281(0.057, 0.506)**	0.389(-0.030, 0.809)	-0.273(-0.590,0.024)
22	-0.557(-1.323, 0.258)	0.188(-0.062, 0.447)	0.302(-0.249, 0.839)	-0.005(-0.250,0.240)
23	-0.1894(-0.631, 0.258)	-0.031(-0.244, 0.190)	0.256(-0.102, 0.605)	**0.341(0.050,0.635)**
24	0.08468(-0.414, 0.648)	0.014(-0.193, 0.226)	-0.269(-0.843, 0.229)	0.014(-0.262,0.287)
25	0.2568(-0.160, 0.703)	**0.346(0.034, 0.634)**	-0.062(-0.407, 0.254)	-0.099(-0.449,0.248)
26	**-0.8051(-1.149, -0.464)**	-0.165(-0.432, 0.098)	**0.538(0.239, 0.839)**	-0.073(-0.371,0.228)
27	**0.6653(0.322, 1.048)**	0.276(-0.005, 0.563)	**-0.486(-0.804, -0.195)**	-0.047(-0.351,0.257)
28	**0.5117(0.158, 0.904)**	0.037(-0.268, 0.332)	**-0.401(-0.700, -0.117)**	0.076(-0.169,0.328)
29	**1.324(0.803, 1.917)**	0.032(-0.217, 0.291)	**-0.753(-1.250, -0.358)**	0.021(-0.216,0.254)
30	**-0.5589(-0.944, -0.193)**	-0.111(-0.336, 0.113)	0.229(-0.084, 0.542)	-0.462(-1.060,0.087)
31	-0.3014(-0.801, 0.215)	0.132(-0.128, 0.378)	0.251(-0.157, 0.652)	0.186(-0.293,0.633)
32	-0.0253(-0.580, 0.583)	0.124(-0.108, 0.353)	0.132(-0.247, 0.533)	-0.112(-0.668,0.415)
33	-0.1393(-0.575, 0.345)	-0.021(-0.230, 0.187)	0.018(-0.342, 0.353)	0.089(-0.507,0.686)
34	**-0.6174(-0.955, -0.278)**	-0.061(-0.306, 0.185)	**0.522(0.227, 0.840)**	0.229(-0.045,0.503)
35	**2.043(1.402, 2.826)**	0.397(-0.512, 1.331)	**-1.289(-1.787, -0.859)**	0.104(-1.039,1.219)
36	-0.1521(-0.470, 0.173)	0.080(-0.223, 0.370)	0.012(-0.255, 0.275)	-0.118(-1.051,0.850)
37	**-0.554(-0.898, -0.226)**	**-0.455(-0.765, -0.147)**	**0.473(0.197, 0.764)**	-0.004(-0.485,0.507)
38	**1.041(0.600, 1.544)**	-0.258(-1.264, 0.691)	**-0.744(-1.074, -0.438)**	-0.060(-1.264,1.140)
Number of DIF Items	13	7	13	2

Table 4.19: Magnitude of Student-Level DIF Values for Mathematics Section: With Covariates for $K = 1$

Item	$G = 1 vs. G = 2$	$G = 1 vs. G = 3$	$G = 1 vs. G = 4$	$G = 2 vs. G = 3$	$G = 2 vs. G = 4$	$G = 3 vs. G = 4$
1	-0.057(-0.499, 0.437)	0.389(-0.408, 1.134)	-0.165(-0.584, 0.298)	0.447(-0.136, 0.964)	0.107(-0.141, 0.361)	0.554(-0.005, 1.047)
2	0.235(-0.178, 0.705)	0.292(-0.440, 0.999)	-0.202(-0.584, 0.224)	0.057(-0.498, 0.554)	**0.436(0.183, 0.685)**	0.494(-0.028, 0.965)
3	**1.499(0.999, 2.101)**	-0.018(-1.027, 0.993)	**2.419(1.902, 3.039)**	**-1.517(-2.264, -0.873)**	**-0.920(-1.189, -0.662)**	**-2.437(-3.203, -1.797)**
4	**0.545(0.212, 0.905)**	**0.681(0.161, 1.214)**	**0.924(0.609, 1.269)**	0.135(-0.260, 0.537)	**-0.379(-0.615, -0.152)**	-0.243(-0.607, 0.136)
5	-0.093(-0.475, 0.322)	0.592(-0.034, 1.216)	0.185(-0.159, 0.558)	**0.685(0.244, 1.128)**	**-0.278(-0.511, -0.055)**	0.407(-0.028, 0.830)
6	**0.657(0.322, 1.029)**	0.203(-0.410, 0.815)	**1.099(0.780, 1.453)**	**-0.454(-0.912, -0.012)**	**-0.442(-0.670, -0.215)**	**-0.896(-1.348, -0.463)**
7	**-0.852(-1.156, -0.538)**	-0.067(-0.587, 0.446)	-0.162(-0.448, 0.130)	**0.785(0.373, 1.170)**	**-0.690(-0.940, -0.436)**	0.095(-0.305, 0.466)
8	0.123(-0.181, 0.441)	-0.172(-0.760, 0.352)	0.142(-0.144, 0.436)	-0.294(-0.760, 0.099)	-0.020(-0.267, 0.223)	-0.314(-0.749, 0.076)
9	**0.525(0.214, 0.848)**	-0.099(-0.682, 0.435)	**0.961(0.656, 1.278)**	**-0.624(-1.075, -0.225)**	**-0.437(-0.691, -0.190)**	**-1.060(-1.523, -0.661)**
10	-0.145(-0.412, 0.136)	0.276(-0.156, 0.721)	0.009(-0.260, 0.286)	**0.421(0.089, 0.750)**	-0.154(-0.384, 0.068)	0.267(-0.055, 0.600)
11	**-0.518(-0.770, -0.262)**	**-0.633(-1.118, -0.198)**	**-0.967(-1.220, -0.713)**	-0.115(-0.538, 0.240)	**0.449(0.228, 0.668)**	0.333(-0.075, 0.678)
12	**0.287(0.006, 0.580)**	0.036(-0.412, 0.472)	0.049(-0.221, 0.323)	-0.250(-0.619, 0.093)	0.238(-0.016, 0.506)	-0.012(-0.356, 0.314)
13	0.127(-0.150, 0.403)	-0.222(-0.671, 0.192)	**-0.339(-0.605, -0.082)**	**-0.348(-0.710, -0.012)**	**0.465(0.225, 0.702)**	0.117(-0.235, 0.443)
14	-0.225(-0.487, 0.041)	0.084(-0.270, 0.450)	**-0.456(-0.702, -0.205)**	**0.309(0.028, 0.602)**	0.231(-0.020, 0.479)	**0.540(0.239, 0.841)**
15	**0.944(0.626, 1.303)**	0.247(-0.125, 0.644)	**-0.466(-0.744, -0.179)**	**-0.697(-1.027, -0.353)**	**1.410(1.105, 1.731)**	**0.713(0.414, 1.011)**
16	**-1.319(-1.739, -0.911)**	**-0.640(-1.118, -0.179)**	**-2.507(-2.866, -2.170)**	**0.680(0.292, 1.013)**	**1.188(0.881, 1.497)**	**1.868(1.518, 2.179)**
17	**-0.606(-0.921, -0.293)**	**0.640(0.250, 1.000)**	**-1.309(-1.595, -1.026)**	**1.246(0.940, 1.556)**	**0.703(0.424, 0.982)**	**1.949(1.653, 2.252)**
18	**-1.526(-1.910, -1.150)**	0.160(-0.388, 0.616)	**-2.397(-2.763, -2.042)**	**1.685(1.343, 1.982)**	**0.872(0.593, 1.141)**	**2.557(2.229, 2.859)**
19	**-0.542(-0.942, -0.134)**	0.066(-0.365, 0.466)	**-2.405(-2.722, -2.087)**	**0.608(0.231, 0.967)**	**1.863(1.521, 2.220)**	**2.471(2.160, 2.775)**
20	**-0.635(-1.013, -0.254)**	-0.251(-0.762, 0.173)	**-1.678(-2.031, -1.343)**	0.384(-0.023, 0.718)	**1.044(0.719, 1.370)**	**1.428(1.064, 1.752)**
21	0.267(-0.066, 0.620)	0.357(-0.159, 0.889)	**1.220(0.908, 1.557)**	0.089(-0.300, 0.486)	**-0.953(-1.200, -0.703)**	**-0.864(-1.259, -0.463)**
22	**0.558(0.107, 1.081)**	0.576(-0.162, 1.336)	**1.662(1.242, 2.161)**	0.018(-0.502, 0.519)	**-1.104(-1.365, -0.844)**	**-1.086(-1.596, -0.616)**
23	**0.583(0.300, 0.880)**	0.165(-0.287, 0.592)	**1.009(0.716, 1.317)**	**-0.418(-0.780, -0.075)**	**-0.426(-0.700, -0.148)**	**-0.844(-1.217, -0.491)**
24	**0.386(0.104, 0.680)**	-0.569(-1.194, 0.033)	**0.940(0.648, 1.258)**	**-0.955(-1.493, -0.473)**	**-0.554(-0.810, -0.305)**	**-1.509(-2.053, -1.028)**
25	0.307(-0.007, 0.615)	-0.159(-0.589, 0.247)	**0.332(0.051, 0.622)**	**-0.466(-0.854, -0.111)**	-0.025(-0.347, 0.318)	**-0.491(-0.858, -0.146)**
26	**-0.750(-1.045, -0.448)**	**0.648(0.286, 1.010)**	**-1.643(-1.926, -1.355)**	**1.397(1.103, 1.700)**	**0.893(0.621, 1.174)**	**2.291(1.998, 2.597)**
27	**-0.332(-0.657, -0.012)**	**-0.527(-0.933, -0.145)**	**-1.353(-1.654, -1.071)**	-0.195(-0.570, 0.128)	**1.021(0.728, 1.317)**	**0.827(0.484, 1.133)**
28	**-1.178(-1.565, -0.805)**	**-0.584(-1.003, -0.189)**	**-2.384(-2.715, -2.077)**	**0.594(0.254, 0.898)**	**1.206(0.941, 1.473)**	**1.800(1.508, 2.072)**
29	**-0.577(-0.851, -0.302)**	**-1.031(-1.558, -0.573)**	**-1.142(-1.415, -0.878)**	**-0.455(-0.967, -0.068)**	**0.565(0.327, 0.794)**	0.110(-0.378, 0.487)
30	0.163(-0.122, 0.463)	0.208(-0.201, 0.609)	**1.381(1.017, 1.760)**	0.045(-0.265, 0.360)	**-1.218(-1.591, -0.878)**	**-1.173(-1.596, -0.782)**
31	0.130(-0.203, 0.482)	0.159(-0.361, 0.677)	**1.548(1.173, 1.954)**	0.028(-0.368, 0.416)	**-1.418(-1.726, -1.131)**	**-1.390(-1.842, -0.965)**
32	**1.028(0.702, 1.377)**	0.414(-0.096, 0.949)	**2.448(2.047, 2.888)**	**-0.614(-1.000, -0.208)**	**-1.419(-1.809, -1.058)**	**-2.034(-2.514, -1.564)**
33	0.244(-0.029, 0.536)	0.092(-0.356, 0.532)	**1.749(1.343, 2.216)**	-0.152(-0.503, 0.181)	**-1.505(-1.939, -1.114)**	**-1.656(-2.169, -1.197)**
34	**-0.771(-1.068, -0.478)**	**0.650(0.298, 0.990)**	**-1.292(-1.580, -0.992)**	**1.421(1.142, 1.717)**	**0.521(0.245, 0.785)**	**1.942(1.653, 2.252)**
35	0.820(-0.131, 1.791)	**-1.716(-2.426, -1.144)**	0.518(-0.449, 1.469)	**-2.537(-3.364, -1.856)**	0.302(-0.763, 1.368)	**-2.235(-3.091, -1.477)**
36	-0.296(-0.617, 0.027)	0.031(-0.347, 0.391)	**1.159(0.534, 1.914)**	**0.328(0.014, 0.633)**	**-1.455(-2.219, -0.813)**	**-1.127(-1.836, -0.517)**
37	-0.490(-0.821, 0.150)	**0.583(0.211, 0.948)**	-0.305(-0.691, 0.093)	**1.073(0.780, 1.367)**	-0.185(-0.584, 0.189)	**0.888(0.503, 1.253)**
38	**1.486(0.761, 2.265)**	**-0.861(-1.360, -0.420)**	**1.419(0.583, 2.362)**	**-2.346(-3.059, -1.734)**	0.067(-0.993, 1.112)	**-2.280(-3.163, -1.507)**
Number of DIF Items	23	12	29	25	29	27

Table 4.20: Magnitude of Student-Level DIF Values for Mathematics Section: With Covariates for $K = 2$

Item	$G = 1 Vs. G = 2$	$G = 1 Vs. G = 3$	$G = 1 Vs. G = 4$	$G = 2 Vs. G = 3$	$G = 2 Vs. G = 4$	$G = 3 Vs. G = 4$
1	-0.619(-1.223, 0.095)	-0.444(-1.219, 0.3923)	-0.595(-1.208, 0.125)	0.175(-0.233, 0.554)	-0.023(-0.283, 0.232)	0.151(-0.216, 0.493)
2	0.307(-0.344, 1.066)	0.08782(-0.6709, 0.9507)	-0.425(-1.082, 0.358)	-0.219(-0.559, 0.102)	**0.732(0.474, 0.984)**	**0.513(0.202, 0.809)**
3	**1.201(0.423, 2.131)**	-0.3002(-1.322, 0.8065)	**1.884(1.123, 2.828)**	**-1.502(-2.032, -1.059)**	**-0.682(-0.944, -0.424)**	**-2.184(-2.714, -1.738)**
4	-0.313(-0.740, 0.111)	**-0.5772(-1.053, -0.06691)**	0.209(-0.230, 0.663)	-0.265(-0.546, 0.010)	**-0.521(-0.797, -0.261)**	**-0.786(-1.055, -0.521)**
5	-0.483(-1.003, 0.084)	-0.5149(-1.155, 0.15)	-0.275(-0.807, 0.304)	-0.032(-0.368, 0.286)	-0.208(-0.450, 0.022)	-0.240(-0.539, 0.047)
6	0.246(-0.301, 0.894)	0.06346(-0.5784, 0.7795)	**1.164(0.593, 1.834)**	-0.182(-0.465, 0.085)	**-0.918(-1.184, -0.670)**	**-1.100(-1.400, -0.815)**
7	**-1.020(-1.460, -0.554)**	-0.307(-0.8074, 0.2227)	-0.165(-0.614, 0.309)	**0.713(0.447, 0.977)**	**-0.855(-1.100, -0.609)**	-0.142(-0.405, 0.110)
8	-0.126(-0.571, 0.327)	**-0.5831(-1.116, -0.03229)**	0.052(-0.404, 0.528)	**-0.458(-0.763, -0.161)**	-0.178(-0.423, 0.076)	**-0.635(-0.925, -0.364)**
9	0.405(-0.046, 0.931)	-0.2551(-0.8098, 0.3584)	**0.619(0.134, 1.157)**	**-0.660(-0.962, -0.383)**	-0.214(-0.476, 0.050)	**-0.874(-1.157, -0.605)**
10	-0.065(-0.459, 0.354)	-0.1378(-0.5796, 0.3395)	-0.049(-0.462, 0.382)	-0.073(-0.317, 0.164)	-0.016(-0.272, 0.236)	-0.089(-0.330, 0.150)
11	**0.533(0.054, 1.075)**	**0.7134(0.1923, 1.309)**	0.115(-0.375, 0.687)	0.180(-0.052, 0.405)	**0.418(0.177, 0.652)**	**0.598(0.362, 0.827)**
12	-0.102(-0.455, 0.277)	-0.2767(-0.6758, 0.1403)	-0.260(-0.629, 0.123)	-0.175(-0.428, 0.073)	0.158(-0.109, 0.436)	-0.017(-0.262, 0.224)
13	**0.775(0.369, 1.231)**	**0.5879(0.1333, 1.08)**	0.237(-0.187, 0.709)	-0.187(-0.429, 0.058)	**0.539(0.276, 0.791)**	**0.351(0.112, 0.601)**
14	**-0.417(-0.738, -0.082)**	-0.1218(-0.4934, 0.2755)	**-0.606(-0.966, -0.237)**	**0.296(0.059, 0.530)**	0.188(-0.091, 0.451)	**0.484(0.236, 0.731)**
15	**0.808(0.443, 1.197)**	0.3174(-0.08844, 0.7544)	**-0.457(-0.835, -0.032)**	**-0.491(-0.775, -0.201)**	**1.265(0.974, 1.549)**	**0.775(0.536, 1.014)**
16	-0.227(-0.559, 0.127)	**0.572(0.2275, 0.9393)**	**-1.960(-2.264, -1.632)**	**0.799(0.435, 1.154)**	**1.733(1.411, 2.056)**	**2.532(2.244, 2.837)**
17	**-0.936(-1.248, -0.623)**	-0.2014(-0.5414, 0.1644)	**-2.115(-2.425, -1.799)**	**0.734(0.442, 1.029)**	**1.179(0.887, 1.465)**	**1.913(1.651, 2.177)**
18	**-1.447(-1.754, -1.139)**	-0.1499(-0.5374, 0.2485)	**-2.561(-2.867, -2.235)**	**1.297(0.993, 1.608)**	**1.114(0.838, 1.385)**	**2.411(2.116, 2.700)**
19	**-0.786(-1.116, -0.447)**	0.3011(-0.09872, 0.7153)	**-2.333(-2.664, -1.998)**	**1.087(0.718, 1.456)**	**1.548(1.227, 1.874)**	**2.635(2.310, 2.975)**
20	-0.238(-0.590, 0.124)	0.181(-0.1711, 0.5451)	**-1.667(-2.007, -1.324)**	**0.419(0.072, 0.762)**	**1.429(1.084, 1.784)**	**1.848(1.554, 2.138)**
21	-0.312(-0.781, 0.204)	-0.3314(-0.8857, 0.2786)	**1.194(0.672, 1.765)**	-0.019(-0.314, 0.264)	**-1.506(-1.805, -1.231)**	**-1.525(-1.860, -1.214)**
22	-0.187(-0.808, 0.524)	-0.2832(-1.034, 0.5516)	**1.110(0.458, 1.834)**	-0.096(-0.456, 0.251)	**-1.297(-1.562, -1.033)**	**-1.393(-1.738, -1.078)**
23	**0.425(0.034, 0.836)**	-0.2807(-0.7452, 0.2108)	**0.479(0.036, 0.918)**	**-0.705(-0.963, -0.450)**	-0.054(-0.331, 0.227)	**-0.759(-1.033, -0.488)**
24	0.456(-0.046, 0.988)	-0.2149(-0.8308, 0.4165)	**1.011(0.496, 1.565)**	**-0.671(-0.960, -0.390)**	**-0.555(-0.840, -0.291)**	**-1.226(-1.544, -0.931)**
25	0.218(-0.172, 0.650)	0.1599(-0.2636, 0.6203)	**0.688(0.255, 1.145)**	-0.058(-0.314, 0.184)	**-0.470(-0.791, -0.149)**	**-0.528(-0.815, -0.254)**
26	**-1.390(-1.687, -1.092)**	**-0.6952(-1.024, -0.3483)**	**-2.375(-2.701, -2.059)**	**0.694(0.417, 0.967)**	**0.985(0.701, 1.281)**	**1.680(1.407, 1.954)**
27	0.057(-0.272, 0.415)	**0.624(0.2758, 1)**	**-0.641(-0.989, -0.270)**	**0.567(0.285, 0.848)**	**0.698(0.397, 0.998)**	**1.265(0.992, 1.542)**
28	**-0.703(-1.021, -0.365)**	0.3285(-0.02277, 0.7026)	**-1.949(-2.258, -1.626)**	**1.031(0.747, 1.308)**	**1.246(0.976, 1.526)**	**2.277(2.029, 2.534)**
29	**0.715(0.208, 1.300)**	**1.046(0.5053, 1.661)**	0.162(-0.340, 0.743)	**0.331(0.086, 0.584)**	**0.554(0.296, 0.799)**	**0.885(0.647, 1.123)**
30	-0.285(-0.640, 0.051)	**-0.5793(-0.9648, -0.196)**	**1.284(0.747, 1.873)**	**-0.295(-0.551, -0.050)**	**-1.569(-2.128, -1.079)**	**-1.864(-2.382, -1.407)**
31	-0.303(-0.747, 0.167)	-0.3937(-0.909, 0.1438)	**1.061(0.553, 1.602)**	-0.090(-0.378, 0.185)	**-1.364(-1.693, -1.071)**	**-1.455(-1.843, -1.112)**
32	**0.879(0.393, 1.444)**	0.2563(-0.2973, 0.876)	**2.535(1.948, 3.235)**	**-0.622(-0.877, -0.371)**	**-1.656(-2.128, -1.241)**	**-2.278(-2.743, -1.874)**
33	0.126(-0.270, 0.587)	-0.06463(-0.5343, 0.4781)	**1.520(0.972, 2.181)**	-0.190(-0.437, 0.057)	**-1.395(-1.875, -0.987)**	**-1.585(-2.050, -1.185)**
34	**-1.328(-1.620, -1.034)**	**-0.4896(-0.819, -0.1322)**	**-2.138(-2.451, -1.824)**	**0.838(0.568, 1.107)**	**0.810(0.534, 1.085)**	**1.648(1.381, 1.919)**
35	**2.466(1.840, 3.185)**	**1.615(1.162, 2.105)**	**2.457(1.643, 3.397)**	**-0.851(-1.563, -0.162)**	0.008(-1.017, 1.001)	**-0.842(-1.761, -0.008)**
36	**-0.528(-0.830, -0.220)**	-0.133(-0.4538, 0.1955)	**1.125(0.482, 1.843)**	**0.395(0.124, 0.664)**	**-1.653(-2.394, -0.992)**	**-1.258(-1.949, -0.638)**
37	**-0.589(-0.906, -0.266)**	**-0.4438(-0.7683, -0.112)**	**-0.855(-1.288, -0.426)**	0.146(-0.177, 0.463)	0.265(-0.190, 0.725)	**0.411(0.005, 0.807)**
38	**2.785(2.083, 3.638)**	**0.9247(0.5395, 1.305)**	**2.521(1.644, 3.491)**	**-1.860(-2.725, -1.148)**	0.264(-0.884, 1.420)	**-1.596(-2.560, -0.734)**
Number of DIF Items	19	13	27	23	27	33

Student- and School-Level Characteristics of Latent Classes. Tables 4.22 and 4.23 present the association analysis result between the estimated student-level group membership and manifest group membership at both student-level and school-level. As shown in Table 4.22, there were significant associations between the estimated group membership and ethnicity and the estimated group membership and gender. Class 1 is predominantly White, Class 2 is predominantly Mexican and Hispanic, Class 3 is predominantly Asian, and Class 4 is predominantly African American. At the school-level shown in Table 4.23, there were significant associations between the estimated school-level group membership and Title I schoolwide program, household income, and poverty-level code, respectively. 88% of those in a Title I schoolwide program were categorized into $K = 1$. As the school-level household income increases, schools were more likely to belong to $K = 2$. In addition, all schools having more than 30% poverty level were included into $K = 1$.

Tables 4.24 and 4.25 show student- and school-level covariate effects based on the multinomial logistic regression covariate model. Values in Tables 4.24 and 4.25 are estimated regression coefficients. As shown in Table 4.24, males were more likely than female to belong to Class 1. American Indians were more likely than Whites to belong to Class 4, Asian Americans were less likely than Whites to belong to Class 2, African Americans and Mexican Americans were more likely than Whites to belong to Classes 2 and 4, and Puerto Ricans and Hispanics were more likely than Whites to belong to Class 2. In Table 4.25, only household income and poverty levels (the higher the value is, the higher the poverty level is) were significant among school-level covariates. Schools which have higher household incomes and lower poverty levels were more likely to belong to school Class 2.

Response Patterns in Latent Classes. Mixture portions of MMixIRTM are intended to differentiate among latent groups of examinees with similar patterns of responses. The methods of identifying DIF typically focus on the correct option. Omitted responses can also

provide useful information concerning differential functioning and are sometimes considered useful indicators of speededness (Mroch & Bolt, 2006).

The last five items on the test were examined as a group to determine whether particular response patterns might emerge conditional on group membership using models with and without covariates. These five items were gridded-in items and were classified as "high" level difficulty in the item descriptions provided by The College Board. There were some noticeable patterns of omitted responses: 99900, 99909, 99990, and 99999 (where 0, 1, and 9 indicate incorrect, correct, and omitted responses, respectively). No students assigned to Class 3, the high ability group, had any omitted responses. Students with 0s and 1s, that is, students who tried to answer the question, were mostly classified into the average and high ability groups. Table 4.26 shows this pattern for the model without covariates. There was a similar pattern to that shown Table 4.26 for the model with covariates.

Table 4.21: Q-Matrix of Mathematics Section

Item	Skill 1	Skill 2	Skill 3	Skill 4	Skill 5	Skill 6	Skill 7	Skill 8	Skill 9	Skill 10	Skill 11
1	1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	1	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0
4	1	0	1	0	0	1	0	0	0	0	0
5	0	0	1	0	0	0	0	0	0	1	1
6	0	1	0	1	0	0	0	0	0	0	0
7	0	0	0	0	1	0	0	0	1	0	0
8	1	0	0	0	0	0	0	0	0	0	0
9	0	1	0	0	0	0	0	0	0	0	0
10	0	0	0	1	1	0	0	0	0	1	1
11	0	0	0	0	1	0	0	0	1	0	0
12	1	0	0	0	1	1	0	0	0	0	0
13	0	0	0	0	1	0	0	0	0	0	0
14	1	0	0	0	1	1	0	0	0	1	1
15	0	0	1	1	1	1	1	0	0	0	0
16	1	0	0	0	1	1	0	1	0	1	1
17	0	0	0	0	1	0	0	0	1	1	1
18	0	1	0	1	1	1	1	0	0	0	0
19	0	0	0	0	1	0	0	1	0	1	1
20	0	1	0	0	1	1	0	0	1	0	1
21	0	0	0	0	1	0	0	0	0	0	0
22	0	0	1	0	0	0	0	0	0	1	0
23	0	0	0	0	1	0	0	0	0	0	0
24	1	0	0	0	0	0	0	1	0	0	1
25	0	1	0	0	0	0	0	0	0	0	0
26	1	0	0	0	0	0	0	1	0	0	0
27	1	0	0	1	1	1	0	0	0	0	0
28	0	1	0	0	0	0	0	0	1	0	0
29	0	0	0	0	1	0	0	0	0	0	0
30	0	0	0	1	1	0	0	0	0	0	0
31	1	0	0	0	0	0	1	0	0	0	0
32	1	1	0	1	0	1	1	0	0	1	0
33	0	0	1	0	0	0	0	0	0	1	0
34	0	1	0	1	0	0	0	1	0	1	0
35	0	0	0	0	1	0	1	0	0	0	0
36	1	1	0	1	0	1	0	0	1	1	0
37	1	0	0	0	0	0	0	1	0	1	0
38	1	0	0	0	0	0	1	0	0	0	0

Table 4.22: Chi-Squares Between Group Membership and Demographic Variables for Mathematics Section: Student-Level

		G=1	G=2	G=3	G=4	χ^2	*df*	*p*-value
Gender	Missing	0(.0)	3(.300)	.2(.200)	5(.500)	35.344	6	.000
	Female	1267(.294)	1442(.334)	501(.116)	1103(.256)			
	Male	1272(.315)	1271(.315)	592(.147)	904(.224)			
Ethnicity	No Response	52(.185)	86(.306)	18(.064)	125(.445)	1501.166	24	.000
	American Indian or Alaska Native	13(.245)	17(.321)	7(.132)	16(.302)			
	Asian, Asian-American, or Pacific Islander	174(.316)	130(.236)	184(.335)	62(.113)			
	Black, or African-American	258(.157)	596(.363)	60(.037)	726(.443)			
	Mexican, or Mexican-American	163(.218)	309(.414)	23(.031)	251(.336)			
	Puerto Rican	25(.170)	56(.381)	8(.054)	58(.295)			
	Other Hispanic, Latino, or Latin American	126(.215)	236(.403)	26(.044)	198(.338)			
	White	1653(.404)	1188(.291)	736(.18)	512(.125)			
	Other	75(.278)	98(.363)	33(.122)	64(.237)			

Numbers in columns are frequencies
Values in parentheses are within-category proportions

Table 4.23: Chi-Squares Between Group Membership and Demographic Variables for Mathematics Section: School-Level

		K=1	K=2	χ^2	*df*	*p*-value
Metro Code	Unclassified	2(.500)	2(.500)	6.571	3	.087
	Rural	13(.520)	12(.480)			
	Suburban	42(.393)	65(.607)			
	Urban	41(.586)	29(.414)			
School Enrollment Size Code	D (300 - 499)	1(1.000)	0(0.000)	1.635	3	.651
	E (500 - 999)	6(.545)	5(.455)			
	F (1,000 - 2,499)	71(.480)	77(.520)			
	G (2,500 - more)	20(.435)	26(.565)			
Title I Schoolwide Program	No/Unknown	77(.423)	105(.577)	17.363	1	.000
	Yes	21(.875)	3(.125)			
Household Income	A ($ 1 - 19,999)	1(1.000)	0(.000)	62.667	16	.000
	B (20,000 - 22,999)	2(1.000)	0(.000)			
	C (23,000 - 24,999)	3(1.000)	0(.000)			
	D (25,000 - 27,999)	6(.667)	3(.333)			
	E (28,000 - 29,999)	2(1.000)	0(.000)			
	F (30,000 - 32,999)	5(.833)	1(.167)			
	G (33,000 - 34,999)	0(.000)	2(1.000)			
	H (35,000 - 37,999)	13(.929)	1(.071)			
	I (38,000 - 39,999)	3(.600)	2(.400)			
	J (40,000 - 42,999)	10(.769)	3(.231)			
	K (43,000 - 44,999)	4(.571)	3(.429)			
	L (45,000 - 47,999)	10(.556)	8(.444)			
	M (48,000 - 49,999)	4(1.000)	0(0.000)			
	N (50,000 - 59,999)	12(.400)	18(.600)			
	O (60,000 - 69,999)	10(.476)	11(.524)			
	P (70,000 Plus)	12(.176)	56(.824)			
	Z (Unclassified)	1(.000)	0(0.000)			
Poverty Level Code	A (0 - 5.9 %)	4(.095)	38(.905)	45.057	3	0.000
	B (6 - 15.9 %)	44(.458)	52(.542)			
	C (16 - 29.9 %)	43(.705)	18(.295)			
	D (30 % - More)	7(1.000)	0(.000)			

Values in parentheses are within-category proportions

Table 4.24: Covariate Effects for Mathematics Section: Student-Level

	$\hat{\gamma_{0gk}}$	$\hat{\gamma_{1g}}(Female)$	$\hat{\gamma_{2g}}(Indian)$	$\hat{\gamma_{3g}}(Asian)$	$\hat{\gamma_{4g}}(Black)$	$\hat{\gamma_{5g}}(Mexican)$	$\hat{\gamma_{6g}}(Puerto)$	$\hat{\gamma_{7g}}(Hispanic)$	$\hat{\gamma_{8g}}(Other)$
G=1	0.359 for $K=1$ -0.573** for $K=2$	-0.247**	-0.438	-0.295	-0.156	-0.809	-0.532	-0.446	-0.116
G=2	0.108 for $K=1$ -0.349** for $K=2$	0.131	0.373	-0.737**	1.492**	1.351**	1.135**	1.238**	0.430
G=3	0* for $K=1$ 0* for $K=2$	0*	0*	0*	0*	0*	0*	0*	0*
G=4	-0.633** for $K=1$ -1.378** for $K=2$	0.140	1.04**	-0.331	2.614**	1.895**	2.149	-0.446	1.088**

* Fixed for model identification

** Significant at $p < .05$

Table 4.25: Covariate Effects for Mathematics Section: School-Level

	$\hat{\gamma_{0k}}$	$\hat{\gamma_{1k}}(Suburban)$	$\hat{\gamma_{2g}}(Rural)$	$\hat{\gamma_{3g}}(SchoolSize)$	$\hat{\gamma_{4g}}(NoTitleIProgram)$	$\hat{\gamma_{5g}}(HouseIncome)$	$\hat{\gamma_{6g}}(PovertyCode)$
K=1	0*	0*	0*	0*	0*	0*	0*
K=2	-1.342	-.549	-0.849	0.371	0.730	0.145**	-.805**

* Fixed for identification

** Significant at $p < .05$

Table 4.26: Response Patterns for Each Latent Class for Last 5 Items Of Mathematics Section

Pattern	Class 1 (Average Ability)	Class 2 (Low Ability)	Class 3 (High Ability)	Class 4 (Very Low Ability)	Total
00000	287	546	14	677	1524
00001	42	3	46	0	91
00009	95	104	4	61	264
00010	87	85	3	33	208
00011	33	3	10	0	46
00019	18	20	2	3	43
00090	7	10	0	17	34
00091	0	0	1	0	1
00099	41	57	1	17	116
00100	102	82	7	8	199
00101	18	2	31	0	51
00109	34	21	2	1	58
00110	37	16	3	0	56
00111	14	0	13	0	27
00119	9	6	0	0	15
00190	3	3	1	0	7
00191	1	0	3	0	4
00199	6	5	0	1	12
00900	7	19	0	20	46
00909	13	24	1	15	53
00910	0	2	0	4	6
00911	0	1	0	0	1
00919	2	3	0	1	6
00990	1	4	0	6	11
00999	30	42	2	22	96
01000	16	9	40	1	66
01001	3	0	59	0	62
01009	11	1	21	0	33
01010	13	1	15	0	29
01011	3	0	38	0	41
01019	4	0	8	0	12
01090	2	0	1	0	3
01091	0	0	2	0	2
01099	3	2	10	1	16
01100	9	1	31	0	41
01101	3	0	55	0	58
01109	4	0	16	0	20
01110	5	0	17	0	22
01111	0	0	55	0	55
01119	3	0	6	0	9
01190	2	0	0	0	2
01191	0	0	2	0	2
01199	1	0	4	0	5
01900	1	0	2	0	3
01901	0	0	1	0	1
01909	0	0	1	0	1
01919	1	0	0	0	1
01999	3	0	2	0	5
09000	101	118	3	67	289
09001	35	4	15	0	54
09009	157	211	1	55	424
09010	25	13	4	5	47
09011	9	0	6	0	15
09019	48	32	0	10	90
09090	6	14	1	8	29
09091	2	0	1	0	3
09099	50	50	2	18	120
09100	43	25	3	1	72
09101	13	1	9	0	23
09109	80	55	4	1	140
09110	18	1	6	0	25
09111	7	0	2	0	9
09119	24	6	1	0	31
09190	3	1	0	1	5
09191	1	0	0	0	1

Table 4.26 continued: Response Patterns for Each Latent Class for Last 5 Items Of Mathematics Section

Pattern	Class 1 (Average Ability)	Class 2 (Low Ability)	Class 3 (High Ability)	Class 4 (Very Low Ability)	Total
09199	17	21	4	1	43
09900	9	19	0	26	54
09901	1	0	0	1	2
09909	29	77	0	43	149
09910	4	4	0	1	9
09911	0	0	1	0	1
09919	4	17	0	3	24
09990	5	12	0	18	35
09991	1	0	0	0	1
09999	60	135	2	87	284
10000	80	154	3	156	393
10001	35	1	10	0	46
10009	18	37	2	14	71
10010	43	19	2	14	78
10011	5	0	16	0	21
10019	12	4	1	1	18
10090	0	0	0	8	8
10091	0	1	0	0	1
10099	13	12	2	5	32
10100	36	23	2	0	61
10101	11	0	22	0	33
10109	18	4	0	0	22
10110	22	2	10	0	34
10111	7	0	22	0	29
10119	4	0	1	0	5
10190	0	1	1	0	2
10199	3	5	0	0	8
10900	1	2	0	9	12
10909	2	10	0	9	21
10910	0	1	0	2	3
10919	1	2	0	1	4
10990	1	2	0	2	5
10999	12	10	1	5	28
11000	16	6	19	1	42
11001	2	0	38	0	40
11009	5	1	7	0	13
11010	10	0	9	0	19
11011	2	0	28	0	30
11019	2	0	6	0	8
11099	1	1	3	0	5
11100	11	0	21	0	32
11101	2	0	62	0	64
11109	1	1	6	0	8
11110	6	0	29	0	35
11111	0	0	112	0	112
11119	3	0	4	0	7
11190	0	0	1	0	1
11199	2	0	3	0	5
11909	1	0	0	0	1
11919	0	0	2	0	2
11999	2	1	2	0	5
19000	38	33	0	16	87
19001	14	1	2	0	17
19009	59	62	0	16	137
19010	8	8	0	2	18
19011	9	1	3	0	13
19019	18	15	0	4	37
19090	2	1	0	1	4
19091	1	0	0	0	1
19099	19	14	0	8	41
19100	20	3	3	0	26
19101	13	0	10	0	23
19109	24	23	1	0	48
19110	13	1	1	0	15
19111	4	0	6	0	10
19119	12	4	1	0	17

Table 4.26 continued: Response Patterns for Each Latent Class for Last 5 Items Of Mathematics Section

Pattern	Class 1 (Average Ability)	Class 2 (Low Ability)	Class 3 (High Ability)	Class 4 (Very Low Ability)	Total
19199	9	5	0	0	14
19900	2	5	0	9	16
19901	2	1	0	0	3
19909	11	25	0	15	51
19910	1	0	0	1	2
19911	1	0	0	0	1
19919	0	6	0	3	9
19990	1	3	0	5	9
19991	1	0	0	0	1
19999	20	33	1	21	75
90000	6	5	0	5	16
90001	1	0	1	0	2
90009	9	4	0	4	17
90010	0	1	0	0	1
90019	2	1	0	0	3
90099	4	6	0	1	11
90100	1	0	0	0	1
90101	1	0	0	0	1
90109	1	2	0	0	3
90119	1	0	0	0	1
90900	0	2	0	2	4
90909	0	6	0	3	9
90910	0	0	0	2	2
90919	0	2	0	0	2
90990	0	0	0	1	1
90999	2	5	0	6	13
91000	0	0	2	0	2
91009	2	0	1	0	3
91010	0	0	1	0	1
91101	0	0	5	0	5
91109	0	0	1	0	1
91111	0	0	4	0	4
91119	0	0	1	0	1
91191	0	0	1	0	1
91199	0	0	2	0	2
91910	0	0	1	0	1
91919	1	0	1	0	2
99000	2	4	0	4	10
99001	1	0	0	0	1
99009	12	18	1	13	44
99010	2	2	0	2	6
99011	1	0	0	0	1
99019	6	3	0	0	9
99090	0	1	0	2	3
99091	0	0	1	0	1
99099	8	13	0	7	28
99100	1	0	1	0	2
99101	1	0	0	0	1
99109	7	6	0	0	13
99110	0	1	0	0	1
99119	2	1	0	0	3
99199	4	4	0	0	8
99900	1	4	0	10	15
99901	2	1	1	0	4
99909	6	22	0	32	60
99910	1	0	0	1	2
99919	2	2	0	5	9
99990	0	8	0	15	23
99999	31	160	0	335	526
Total	2539	2716	1095	2012	8362

4.2.3 STD P-DIF Results

One school was selected as the focal group to illustrate the results from the MMixIRTM and the STD P-DIF approaches. In this section, we report the STD P-DIF results. The comparison with the MMixIRTM results is given in the next section. The focal group school was selected because it had the largest number of students taking the PSAT/NMSQT, 114 students, according to STD P-DIF comparison group definitions. Score-levels were collapsed across the 20 to 80 point reporting scale used by The College Board, to yield 10 score intervals as follows: 20 - 30, 31 - 40, 41 - 50, 51 - 60, 61 - 70, and 71 - 80. STD P-DIF values are shown in Table 4.27. As mentioned earlier, College Board policy is to inspect STD P-DIF values between -10% and -5% and between 10% and 5% to ensure that no possible DIF effect is overlooked. There were 10 items having STD P-DIF values in these ranges. Items with STD P-DIF values outside the -10% and 10% range are more unusual and are examined carefully. There was one item with such a value.

4.2.4 Comparisons Between STD P-DIF and MMixIRTM

In this section, we compare results from the STD P-DIF and MMixIRTM at the school-level. Detected DIF items based on STD P-DIF were different across schools while those items based on the MMixIRTM were the same within the same school-level group. For comparison purposes, results for the school reported in Table 4.27 are presented. The school used for the analysis of STD P-DIF also was included in school-level class 2 and used for estimating the MMixIRTM.

For comparison purposes, three cases were identified in Table 4.28. Case I were items that were identified as DIF by STD P-DIF, but not by MMixIRTM, Case II were items identified as DIF by MMixIRTM, but not by STD P-DIF, and Case III were items identified DIF by both STD P-DIF and MMixIRTM.

Item categorization results as shown Table 4.28 are different across models without and with covariates. The number of DIF items in Case II was larger than that of Case I especially

with covariate model, which is mainly because the DIF analysis based on latent groups works to find the nuisance dimension(s) along which the latent classes differ and then separates latent classes based on that dimension (or dimensions)

Table 4.27: STD P-DIF Value

Item	STD P-DIF Value
1	0.56
2	-1.75
3	-0.87
4	4.58
5	0.32
6	2.05
7	-1.97
8	0.73
9	-4.61
10	-3.66
11	**5.36**
12	**7.03**
13	**6.18**
14	0.28
15	**5.93**
16	**6.64**
17	**-5.82**
18	2.97
19	-4.04
20	3.93
21	**9.5**
22	**5.67**
23	-2.01
24	**8.61**
25	-2.69
26	-2.86
27	-4.11
28	1.13
29	**-12.25**
30	0.76
31	**7.89**
32	1.52
33	-1.89
34	-0.94
35	2.72
36	-1.39
37	-2.31
38	-1.5

Table 4.28: Result Comparisons with STD P-DIF and MMixIRTM for a School

Model	Case	DIF Items
Without Covariate Model	Case I	Items 11, 12, 16, 17, 21, and 22
	Case II	Items 4, 6, 10, 14, 19, 25, 26, and 38
	Case III	Items 13, 15, 24, 29, and 31
With Covariate Model	Case I	Items 12, 15, 22, 24, and 31
	Case II	Items 2, 3, 4, 5, 6, 10, 23, 25, 26, 27, 28, 30, 34, 35, 37, and 38
	Case III	Items 13, 16, 17, 21, and 29

Chapter 5

CONCLUSIONS

5.1 Summary and Implications of Results

The purpose of this study was to provide a model that would assist The College Board in describing the comparison groups used in reporting DIF results to schools. The model developed in this study, the MMixIRTM, employs features of an IRT model, an unrestricted LCM, and a multilevel model, the description of the model was described from those three perspectives. Model estimation is often an important concern. Therefore, in this study, a a fully Bayesian method for estimation of the model parameters was described using the freely available software, WinBUGS. A simulation study was done to investigate the performance of the estimation algorithm that was developed to study the behavior of the MMixIRTM for detection of DIF under some typical testing conditions. The results indicated that the generated parameters were recovered very well for the conditions considered. Use of MMixIRTM also was illustrated with the PSAT/NMSQT Mathematics Test. Finally, the results from the MMixIRTM were compared with those from the STD P-DIF, the currently used DIF detection method, on a school selected as an example.

There are several differences between the STD P-DIF and the MMixIRTM approaches to DIF detection. First, the DIF detection using the STD P-DIF is based on manifest groups in which a comparable group is composed (a) of those students who have scored between 20 and 80 for all three sections in PSAT/NMSQT and (b) who were in Grades 10 and 11. The comparable group also does not include any non-standard students. The MMixIRTM, on the other hand, uses a model-based method for detecting latent groups in the data at both the student-level and at the school-level. As a result, DIF information from by the

MMixIRTM is provided about student-level differences in response propensities and school-level factors associated with different proportions of student-level latent classes. That is, the MMixIRTM clusters students into latent classes with respect to their response patterns, and schools are clustered with respect to different proportions of student-level groups. With this type of clustering, the MMixIRTM makes it possible to provide a given school with descriptions of student- and school-level characteristics associated with the given school as well as associated with other schools not in the same latent class as the given school. Student-level characteristics include such things as ability levels, response patterns (including patterns of omissions), particular demographic characteristics that predominate in one or more student- or school-level latent classes, and profiles of item parameter values as well as DIF item information about individual items. At the school-level, results from a MMixIRTM analysis provide each school with descriptions of schools in each school-level latent class, including their own. This description can then be used to provide schools with a framework within which to compare the results of their school with other schools in their latent class and in the other latent classes. As an example, with respect to the school used in the example of Mathematics section in this study, those schools that were classified into the same latent class, (i.e., school-level latent group 2) were characterized by lower Title I enrollment, higher household income, lower poverty levels and a predominance of students in student-level latent class 1. Other information describing individual schools can be similarly used to describe members of each school-level latent class.

Second, STD P-DIF is based on observed scores while MMixIRTM is based on latent score from IRT model. The use of latent instead of observed scores provided the potential for separation of the effects of item difficulty and ability, and also enables the modeling of response measurement error (Fox & Glas, 2001).

Third, STD P-DIF provides DIF information for each school using only a single comparable group. The characteristics of this group are described in terms of information that may not be closely related to why students in a given school performed better than or worse

than other schools in the comparable group. The MMixIRTM, on the other hand, provides DIF information for all schools that are members of the same school-level latent class. Further, there is also the possibility that there may be more than one such comparable group. STD P-DIF currently provides DIF information only at the school-level. The MMixIRTM provides DIF information simultaneously at both school- and student-levels. It is further possible to incorporate information regarding specific item-level skill as shown in Figure 5.1 to explain the meaning of the differences in item functioning among the latent groups. For example, in Figure 5.1, 13 Mathematics items are shown to illustrate simultaneous school- and student-level DIF information. In this figure, the x-axis represents item number (from 13 to 25) and the y-axis represents item difficulty. Highlighted values in the table indicate both student-level and school-level DIF items. As an example, items 15, 19, and 25 were detected as a DIF item at the school-level but all except item 13 were detected as DIF items at the student-level. In this study, item-level skill information was not incorporated into the description of items, but this is clearly possible.

5.2 Limitations

Substantive Usefulness of MMixIRTM

The substantive usefulness of the model is based on the assumption that the resulting classes represent discrete subpopulations and not just statistical artifacts of non-normality that may incidentally exist in the data (Bauer & Curran, 2003). At the present time, previous findings are not available indicating what schools look like based on the latent classes. This kind of information can be developed, however, and will soon lead to a better understanding of how many school-level latent classes might be expected. The resulting student- and school-level mixtures in the data examined in the empirical example were clearly distinguishable in terms of ability level, response patterns (such as omission rate), item difficulty profile, and student and school demographic characteristics. When several factors determine a class, however, finding those factors that cause is often difficult.

Label Switching. The second-type of label switching, that is, label switching among different replications of the same simulation conditions was detected in the simulation study. This type of label switching can be problematic (1) for checking convergence using more than two chains, (2) for comparing student-level latent group membership across school-level latent classes (because the student-level probability of mixtures is modeled for each school-level class), and (3) for comparing results with and without covariates.

The WinBUGS code developed for this study was not designed to prevent the second type of label switching. This type of label switching can be easily detected and taken care of in a simulation study, but it can be a problem in an empirical study. The strategy we used in this study to check for this type of label switching was to investigate item difficulty profiles and ability patterns for each school latent class to see if the representation is similar across school-level mixtures, and to crosstabulate group memberships to find the dominant group.

Limitation of the Simulation Study. We investigated model recovery only for binary responses. The behavior of the MMixIRTM for the polytomous responses and mixed binary and polytomous responses needs to be investigated. It may also be useful to investigate the behavior of the MMixIRTM with respect to different numbers of items and examinees.

In this study, only one covariate was used with two different percentages of overlap between the latent classes and manifest covariate. There were no differences in the parameter recovery with- and without-covariate models. However, there were little differences between the two models in the empirical study, when more than one covariate was used, at both student- and school levels. This suggests that additional simulation studies are needed to examine the effect of covariates on parameter recovery.

Use of Priors on the Probabilities of Mixtures. Three different kinds of priors on the probabilities of mixtures were studied along with their hyperparameters, α_g and α_k. For student-level mixture probabilities, a Dirichlet distribution provides a conjugate prior. A Dirichlet distribution with the Gamma distribution also was studied for the hyperparameters,

$\alpha_g = 1$ and $\alpha_g = 4$. At the school-level, probabilities of mixture were studied using a Dirichlet prior as a conjugate prior, a Dirichlet prior with the Gamma distribution, and a Dirichlet process with stick-breaking prior for the hyperparameters, $\alpha_k = 1$ and $\alpha_k = 4$.

A small simulation study for the priors was done to determine whether these priors worked well under the conditions simulated. The Dirichlet distribution with the Gamma distribution with $\alpha_g = 1$ and $\alpha_k = 1$ for both student-level and school-level performed well in that the models were identified. It is important to note that this result is limited to the conditions studied here. More extensive simulation study of the behavior of these priors on the probabilities of mixture is still needed.

Algorithm Efficiency using WinBUGS. von Davier and Yamamoto (2007) have noted that MCMC typically requires substantial computing time to obtain usable results. The use of multiple starting points is necessary with empirical data, in particular, to determine whether stationarity has been attained (i.e., whether the algorithm has converged). For this reason, the amount of computing required can sometimes be very large. In the case of the example in this study: 90 hours were required for one condition in the simulation study and 121.5 hours for the empirical study, respectively, on a 3.0 GHz computer with 1GB of RAM to complete a single replication. In order to implement this model in an operational situation, a much faster algorithm would be required. Results such as this are not uncommon, and MCMC estimation of model parameters for long tests and very large samples is clearly going to require substantial computing resources or development of speedier algorithms. When compared to relatively speedy computation using software implementing EM algorithms for some simpler IRT models, MCMC does not yet appear to be ready for use in most operational testing programs.

5.3 Discussion

Anchoring and Linking for a MMixIRTM DIF Analysis. In a DIF analysis with a MMixIRTM, each latent group of examinees is evaluated over the same set of test items.

Thus, one can think of every item on the scale as being a potential anchor item to be used in estimating an appropriate set of linking coefficients. This design is similar to the common-item nonequivalent group design except that group membership is latent. Following the rationale for that design, group membership and class-specific item difficulty parameters are estimated simultaneously with metric anchoring done in terms of the ability distribution.

An alternative way of anchoring and linking in MMixIRTM is that we set the anchor item(s) before fitting the MMixIRTM. This would be consistent with Thissen, Steinberg, and Wainer (1993) who suggest that a set of anchor items first be identified that contains no DIF. These anchor items then serve to identify the metric of the latent trait scale, and thus item parameters from the non-anchor items are directly comparable. Implementing this kind of apriori determination of DIF-free items is a potentially useful means of establishing a set of anchor items, but how this should be done in the context of a latent groups DIF analysis is not clear. What would be needed to do this would be a substantive rationale clearly indicating what latent groups one might anticipate, given the test and the sample, and then how these latent groups might be expected to perform on particular items. Lacking that kind of rationale, apriori specification of anchor items that are assumed to be DIF-free would seem premature.

Interpretation of Class-Specific Construct Representation. The ability estimates reflect somewhat different processes across classes. The following questions remain (Embretson & Reise, 2000): Should abilities be interpreted equivalently across classes? In addition, do we need to adjust scores for deficient knowledge states?

Use of Covariates and Predictors for MMixIRTM. An important distinction between covariates and predictors is that covariates are variables that may be used to describe or predict (rather than to define or measure) the latent classes while predictors are variables that may be used to describe or predict the dependent variable (Vermunt & Magidson, 2005). In this study, only covariates were used for both simulation and empirical studies. Further study would be useful to explore inclusion of covariates and predictors in a MMixIRTM.

Palardy and Vermunt (2007) note that, to the degree that the predictors adjust the ability estimates, group membership can change. For school comparisons based on student achievement, it is possible to equalize schools in terms of predictors and then determine the number of classes based on achievement. Failing to control for excessive random variation in between-level ability, however, will lead to over extraction of classes.

Type I Error Rate Control of Multiple Pairwise Comparisons. In this study, the HPD interval was used in the detection of DIF items. If more than three latent classes (i.e., $G \geq 3$, $K \geq 3$) were present, then a multigroup DIF analysis is potentially needed (depending on the types of comparisons that are of interest). A concern with multigroup DIF comparisons is control of Type I error.

Two candidate methods for multiple comparisons are the Benjamini-Hochberg (B-H: Steinberg, 2001) sequential approach and Bonferroni methods. Williams, Jones, and Tukey (1999) showed that a sequential approach to controlling the false discovery rate in multiple comparisons yields much greater power than the widely used Bonferroni method. The B-H is a sequential procedure implemented as follows: Consider testing hypotheses $H_1, H_2, \ldots, H_{K-1}$ based on the corresponding p-values, $P_1, P_2, \ldots, P_{K-1}$. Let $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(K-1)}$ be the ordered p-values of the z statistics, and denote by $H_{(r)}$ the null hypothesis corresponding to $P_{(i)}$. Let k be the largest r for which

$$P_{(r)} \leq [\frac{r}{(K-1)+1-r}] \cdot \alpha \tag{5.1}$$

then reject all H_r for $r = 1, 2, \ldots, K-1$.

The above multiple comparisons in the H_0 is the situation where all multiple comparisons are ones with one control group (i.e., the focal group) and planned rather than post-hoc tests. Dunnett's method is recommended for pairwise multiple comparisons with one control group (i.e., the focal group) (Hsu, 1999). In this study, a given school can be a focal group at the school-level and a dominant group can be a focal group at the student-level.

Descriptive Latent DIF Measure. As statistical power is a function of sample size (Cohen, 1988), relatively small differences in population parameters will be statistically sig-

nificant for sufficiently large samples. The result is that we can expect to reject the null hypothesis, in this case of no DIF, when the sample size is very large but the difference between parameters is small, possibly even trivially small (Dorans & Holland, 1993; Kim et al., 2007). Since the College Board often has very large sample sizes for some of its testing programs, a DIF analysis relying solely on statistical tests is not likely to be useful, as most if not all items will be identified as functioning differentially. In addition to testing the significance of differences in item parameter estimates, indices of the magnitude of DIF, that is, effect sizes, often can be used to help determine whether or not the DIF that was detected is sufficiently large to be meaningful. In this study, therefore, we employed indices of the magnitude of DIF suggested for the manifest DIF analysis (e.g., Dorans & Kulick, 1986; Dorans & Schmitt, 1991; Kim, 2000; Wainer, 1993; Zumbo, 1999) to latent DIF analysis after group membership and its class-specific item difficulty were obtained.

5.4 Possible Applications of MMixIRTM and Related Modeling for Educational Research

5.4.1 Applications of MMixIRTM

Educational Policy Research. The MMixIRTM may also have application in school effectiveness research. Most of current school effectiveness research is done to explore differences within and between schools by investigating the relationship between student-level and school-level background variables. The outcome variable is typically a sum score on a test, and the analysis is is usually done by studying differences among schools after adjusting for relevant background variables.

As was shown in the analysis of empirical data, using the MMixIRTM enabled a clustering of students and schools with respect to response patterns. In addition, it was then possible to to determine what may have caused these differences by examining covariates either directly in the model or subsequently, by means of tests of association. In addition, it was possible to show the heterogeneity across groups in the DIF analysis. Results such as these also can

provide useful information for educational policy research mainly by addressing to what homogeneous groups look like based on students' achievement at both the student-level and school-level.

Application MMixIRTM to Longitudinal Data. Longitudinal data can be viewed as a special type of multilevel data in which the first level units refer to responses over time. Typically the second level units are individuals like students. The higher level units can be schools, for example. In a cross-sectional analysis, students are clustered with respect to their response patterns at the first level for cross-sectional data, and growth patterns are clustered with respect to their initial status and growth rate at the first level for longitudinal data. In addition, just as schools are clustered with respect to different proportions of first-level groups (i.e., student-level), at the second level for the cross-sectional data, students are clustered with respect to different proportions of growth patterns for longitudinal data.

5.4.2 Modeling of Other Item Structures

Modeling of Testlet Effect for Reading Test. Standardized educational achievement tests often include sections composed of groups of items based on a common stimulus (e.g., a set of test items all focusing on a single passage in a reading comprehension test). These groups of items, known as testlets, may offer greater efficiency in test construction, as a single stimulus can be associated with several items, but they are also the potential cause of nuisance dimensionality extraneous to the construct being measured. This dimensionality, reflected as local dependence, arises because items within a testlet rely on a common stimulus. Cohen, Cho, and Kim (2005) proposed a MixIRTM for the data structure having testlets. The MMixIRTM should be capable of being extended to incorporate items with a testlet structure.

Non-Uniform DIF Analysis. The MMixIRTM was tested in this study using the Rasch model, which allows the item difficulty to differ across mixtures. The structure of the Rasch model only allows for uniform DIF among latent classes. It should be possible to extend the

MMixIRTM to other models such as the 2- or 3-parameter IRT model to account not only for differences between the groups with respect to item difficulty, but also differences with respect to discriminating power or guessing parameters.

Incorporation of a Q-Matrix in a MMixIRTM for Diagnostic Modeling. In the linear logistic test model (LLTM, Fisher, 1983), cognitive skills are modeled to reflect basic features of items and how they affect probabilities of responses (Tatsuoka, 1983). A Q-matrix (Tatsuoka, 1983) is used to contain this information, the entries of which indicate whether or not a particular cognitive skill is required by attribute h in item i. Elements of the Q-matrix, q_{ih}, are either 1 if attribute h is required by item i or 0 if is is not. The LLTM extends the Rasch model by positing a linear structure for item difficulty, β_i:

$$\beta_i = \sum_h q_{ih}\eta_h = q_{ih}^{'}\eta, \tag{5.2}$$

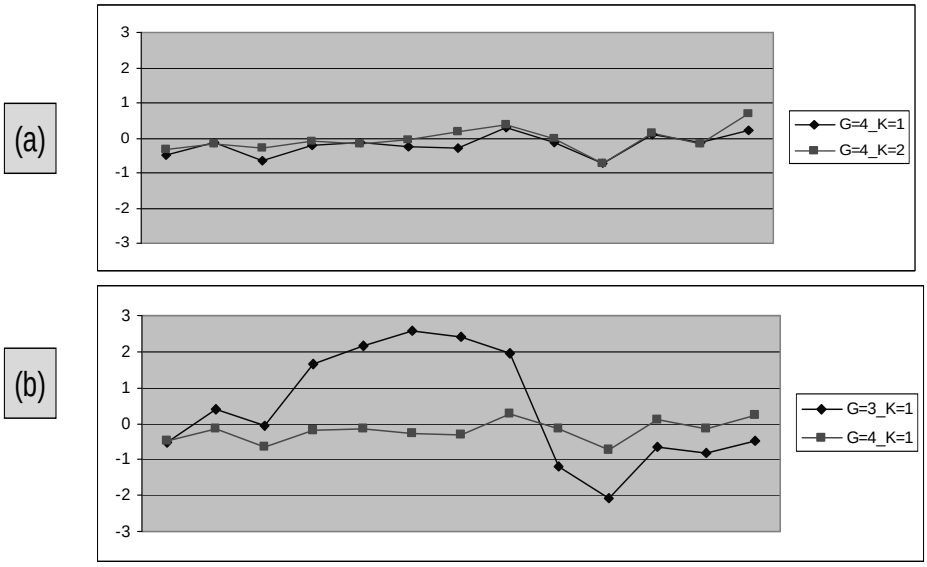
where η_h is a contribution to item difficulty entailed by attribute h.

We can incorporate the following linear structure for the item difficulty in MMixIRTM to support the narrative theme of fundamental measurement:

$$\beta_{igk} = \sum_h q_{igkh}\eta_{gkh} = q_i^{'}\eta_{gk}, \tag{5.3}$$

where η_{gkh} is a contribution to item difficulty required by attribute h for each class, g and k.

Figure 5.1: Item Difficulty Profile with DIF Information and Item Skill Information: (a) For School-Level Comparison $G = 4, K = 1$ Vs. $G = 4, K = 2$ (Selected Classes), and (b) For Student-Level Comparison of $G = 3, K = 1$ Vs. $G = 4, K = 1$ (Selected Classes)



	13	14	15	16	17	18	19	20	21	22	23	24	25
School-Level DIF Magnitute	-.17	.05	-.36	-.10	.04	-.20	-.50	-.11	-.11	.01	-.03	.03	-.47
Student-Level DIF Magnitute	-.04	.53	.56	1.87	2.31	2.87	2.71	1.69	-1.06	-1.35	-.75	-.70	-.71
Skill 1	0	1	0	1	0	0	0	0	0	0	0	1	0
Skill 2	0	0	0	0	0	1	0	1	0	0	0	0	1
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
Skill 11	0	1	0	1	1	0	1	1	0	0	0	1	0

Q-Matrix

Bibliography

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

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

[3] Angoff, W. A. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates.

[4] 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.

[5] Bassiri, D. (1988). *Large and small sample properties of maximum likelihood estimates for the hierarchical linear model.* Unpublished doctoral dissertation, Michigan State University.

[6] Bates, D., & Debroy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis, 91*, 1-17.

[7] Bijmolt, T. H. A., Paas, L, & Vermunt, J. K. (2004). Country and consumer segmentation: Multi-level latent class analysis of financial product ownership. *International Journal of Research in Marketing, 21*, 323-340.

[8] 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.

[9] 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.

[10] Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

[11] Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 3*, 473-514.

[12] Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*, 65-108.

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

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

[15] Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2006, June). *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.

[16] Cohen, A. S., Cho, S.-J., & Kim, S.-H. (2005, April). *A mixture testlet model for educational tests.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

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

[18] 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.

[19] Cohen, J. (1998). *Statistical power analysis for the behavioral sciences* (2nd eds.). Hillsdale, NJ: Erlbaum.

[20] Congdon, P. (2003). *Applied Bayesian modelling.* New York: Wiley.

[21] DeAyala, R. J., Kim, S. -H., Stapleton, L. M., & Dayton, C.M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 2*, 243-276.

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

[23] Diebolt, J., & Robert, C. P. (1994). Estimation of finite distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B, 56*, 363-375.

[24] Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66.) Hillsdale, NJ: Erlbaum.

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

[26] Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (Research Rep. No. RR-91-47). Princeton, NJ: Educational Testing Service.

[27] Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test-third edition.* Circle Pines, MN: American Guidance Service.

[28] Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence-Erlbaum.

[29] Finch, S. J., N. R. Mendell, & H. C. Thode (1989). Probablistic measure of adequacy of a numerical search for global maximum. *Journal of the American Statistical Association, 84*, 1020-1023.

[30] Fisher, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3-26.

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

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

[33] Frühwirth-Schnatter, S. (2006). *Finite mixture and markov switching models.* New York: Springer.

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

[35] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*(2nd Eds). Chapman & Hall.

[36] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4*(pp. 169-194). Oxford: Oxford University Press.

[37] Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for Gibbs' sampling. *Applied Statistician, 41*, 337-348.

[38] Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*, 43-56.

[39] Goldstein, H. (1987). *Multilevel models in education and social research.* London: Charles Griffin and Co.

[40] Goldstein, H. (1989). Restricted unbiased iterative generalized least squares estimation. *Biometrika, 76*, 622-623.

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

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

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

[44] Hu, P. G., & Dorans, N. J.(1989). *The effect of deleting differentially functioning items on equating functions and reported score distributions.* Princeton, NJ: Educational Testing Service.

[45] Johnson, V., & Albert, J. (1998). *Ordinal data modeling.* New York: Springer.

[46] Jordan, M. I., & Jacobs, R. A. (1992). Hierarchies of adaptive experts. In Moody, J., Hanson, S., & Lippmann, R. (Eds.), *Advances in Neural Information Processing Systems, 4* (pp. 985-993), San Mateo, California: Morgan Kaufmann, 985-993.

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

[48] Kamata, A., & Binici, S. (2003, June). *Random-effect DIF analysis via hierarchical generalized linear models.* Paper presented at the International Meeting of the Psychometric Society: The 68st annual meeting of the Psychometric Society, Sardinia, Italy.

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

[50] Kang, T.-H., & Cohen, A. S. (in press). A mixture model analysis of ethnic group DIF. *Journal of Educational Evaluation.*

[51] Kiefer, N. M. & J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidential parameters. *Annals of Mathematical Statistics, 27*, 887-906.

[52] Kim, K.-S. (1990). *Multilevel data analysis: A comparison of analytical alternatives.* Unpublished doctoral dissertation, Los Angeles: University of California.

[53] Kim, S.-H. (April 2000). *An investigation of the likelihood ratio test, the Mantel test, and the generalized MantelHaenszel test of DIF.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

[54] Kim, S.-H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*, 93-116.

[55] Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies.* Unpublished Report. Los Angeles: University of California, Department of Statistics.

[56] Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2006, April). *Model selection methods for mixture dichotomous IRT models.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

[57] Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4*, 115-136.

[58] Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika, 74*, 817-827.

[59] Longford, N. T. (1993). *Random coefficient models.* New York: Oxford University Press.

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

[61] 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.

[62] Maple, S. A., & Stage, F. K. (1991). Influences on the choice of math/science major by gender and ethnicity. *American Educational Research Journal, 28*, 37-60.

[63] Masters, G. N. (1985). A comparison of latent trait and latent class analysis of Likert-type data. *Psychometrika, 49*, 69-82.

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

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

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

[67] Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement, 36*, 217-232.

[68] Muthén, B., & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors, 31*, 1050-1066.

[69] Muthén, B., Brown, C. H., Jo, B. K. M., Khoo, S.-T., Yang, C. C., Wang, C.-P., & Kellam, S. G. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics, 3*, 459-475.

[70] Muthén, L. K. & Muthén, B. O. (2006). Mplus [Computer program]. Los Angeles, CA: Muthén & Muthén.

[71] O'neill, K. A., & Mcpeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum Associates.

[72] Palardy, G., & Vermunt, J. K. (2007). Multilevel growth mixture models for classifying group-level observations. (downloaded from website http://spitswww.uvt.nl/ vermunt/PapersSubmitted)

[73] Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146-178.

[74] Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

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

[76] Pine, S. M. (1977). Applications of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing: Proceedings of a symposium presented a the 18th annual convention of the Military Testing Association* (Research Rep. No. 77-1, pp. 37-43). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

[77] R Development Core Team. (2007). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http: www.R-project.org/.

[78] Raftery, A. L., & Lewis, S. (1992). How many iterations in the Gibbs sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4*(pp. 763-773). Oxford: Oxford University Press.

[79] Rasbash, J., Steele, F., Browne. W., & Prosser, B. (2005). *A user's guide to MLwiN, Version 2.0.* University of Bristol, U.K.

[80] 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.

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

[82] Raudenbush, S. W., Bryk, A. G., & Congdon, R. (2005). HLM: Hierarchical linear and nonlinear modeling [Computer program]. Chicago: Scientific Software International.

[83] Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical*

Society, Series B, 59, 731-792. Correction (1998). *Journal of the Royal Statistical Society Series, Series B, 60*, 661.

[84] Robert, C. P. (1996). Mixtures of distributions: inference and estimation. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 441-464). Washington DC: Chapman & Hall.

[85] Rodríguez, G, & Goodman, N. (1995). An assessment of estimation procedure for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A, 158*, 73-89.

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

[87] 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.

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

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

[90] Schmitt, A.P., Holland, P.W., & Dorans, N.J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281-315). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

[92] Smith, B. (2004, April). Bayesian output analysis program (BOA) Version 1.1.2 for R and S-PLUS [Computer program]. Department of Biostatistics, University of Iowa, Iowa City.

[93] Spiegelhalter, D., Thomas, A., & Best, N. (2003). WinBUGS version 1.4 [Computer program]. Robinson Way, Cambridge CB2 2SR, UK: MRC Biostatistics Unit, Institute of Public Health.

[94] Steinberg, L. (2001). The Consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81*, 332-342.

[95] Stephens, M. (2000). Dealing with Label Switching in Mixture Models. *Journal of the Royal Statistical Society, Series B, 62*, 795-809.

[96] Tate, W. F. (1997). Race-ethnicity, SES, gender, and language proficiency trends in mathematics achievement: An update. *Journal for Research in Mathematics Education, 28*, 652-679.

[97] Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

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

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

[100] Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*, 77-83.

[101] Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology, 33*, 213-239.

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

[103] Vermunt, J. K. (in press a). A hierarchical model for clustering three-way data sets. *Elsevier Science.*

[104] Vermunt, J. K. (in press b). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research.*

[105] Vermunt, J. K. (2007). Multilevel mixture item response theory models: An application in education testing. ISI 2007 Proceedings.

[106] von Davier, M. (2001). WINMIRA [Computer program]. St. Paul, MN: Assessment Systems Corporation.

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

[108] von Davier, M., & Yamamoto, K. (2007). Mixture-Distribution and HYBRID Rasch Models. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 99-115). New York: Springer.

[109] Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123-135). Hillsdale, NJ:Erlbaum.

[110] Webb, M. -y., Cohen, A. S., Schwanenflugel, P. J. (in press). Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test-III. *Educational and Psychological Measurement.*

[111] Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24*, 42-69.

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

[113] Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF).* Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Appendix A

WinBUGS Code Used for Multilevel IRT Model

```
# J: the number of students
# I: the number of items
# T: the number of schools
# b: item difficulty
# eta: ability at the student-level
# nu: ability at the school-level
# gamma: mean of ability at the school-level
# tau (square root of tau in model description): the SD of ability at the student-level
# zeta: inverse variance of ability at the school-Level
model
{
for (j in 1:J) {
    for (i in 1:I) {
    r[j,i]<-resp[j,i]
}}

#  Multilevel Rasch
for (j in 1:J) {
    for (i in 1:I) {
     logit(p[j,i]) <-  tau*eta[j] +  nu[group[j]] - b[i]
     r[j,i]~dbern(p[j,i])
}}


# Ability

for (j in 1:J) {
    eta[j] ~ dnorm(0, 1)
}

for (j in 1:T){
    nu[j] ~ dnorm(gamma, zeta)
}

gamma ~ dnorm(0,1)
```

```
# Standard Deviation of Ability at the Student-Level

tau ~ dnorm(0,1)I(0,)

# (1/Variance) of Ability at the School-Level

zeta ~ dgamma(0.1,0.001)


# Item Difficulty
for (i in 1:I) {
    b[i]~dnorm(0, 1)
 }

# Log Likelihood
for (j in 1:J) {
   for (i in 1:I) {
    l[j,i]<-log(p[j,i])*r[j,i]+log(1-p[j,i])*(1-r[j,i])
 }}
 loglik <-sum(l[1:J,1:I])
 AIC <- -2*(loglik - np)
 BIC <- -2*loglik + np*log(N)
}

list(J=8000, I=40, T=320,  np=42,
group=c(1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1
, 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1
, 1   , 1   , 1   , 1   , 2   , 2   , 2   , 2   , 2   , 2   , 2
, 2   , 2   , 2   , 2   , 2   , 2   , 2   , 2   , 2   , 2   , 2
, 2   , 2   , 2   , 2   , 2   , 2   , 2   , 3   , 3   , 3   , 3
...),
resp=structure(.Data=c(
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,0.0,1.0,0.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,0.0,1.0,1.0,0.0,
...
0.0,1.0,1.0,1.0,0.0,
0.0,0.0,1.0,1.0,0.0,
1.0,1.0,0.0,0.0,0.0,
```

```
1.0,0.0,0.0,0.0,0.0),.Dim = c(8000,40)))
```

Appendix B

WinBUGS Code Used for Rasch Model

```
# J: the number of students
# I: the number of items
# tau (square root of tau in model description): the SD of ability
# b: item difficulty
# eta: ability
model
{
for (j in 1:J) {
    for (k in 1:I) {
    r[j,k]<-resp[j,k]
}}

#  Rasch model
for (j in 1:J) {
    for (i in 1:I) {
     logit(p[j,i]) <-  tau*eta[j]  - b[i]
     r[j,i]~dbern(p[j,i])
}}

# Ability
for (j in 1:J) {
    eta[j] ~ dnorm(0, 1)
}

#  SD of Ability
tau ~ dnorm(0, 1) I(0, )

# Item Difficulty
for (i in 1:I) {
    b[i]~dnorm(0, 1)
 }
# Log Likelihood
for (j in 1:J) {
   for (i in 1:I) {
    l[j,i]<-log(p[j,i])*r[j,i]+log(1-p[j,i])*(1-r[j,i])
 }}
 loglik <-sum(l[1:J,1:I])
```

```
 AIC <- -2*(loglik - np)
 BIC <- -2*loglik + np*log(N)
}

list(J=8000, I=40, np=41,
resp=structure(.Data=c(
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,0.0,1.0,0.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,0.0,1.0,1.0,0.0,
...
0.0,1.0,1.0,1.0,0.0,
0.0,0.0,1.0,1.0,0.0,
1.0,1.0,0.0,0.0,0.0,
1.0,0.0,0.0,0.0,0.0),.Dim = c(8000,40)))
```

Appendix C

WinBUGS Code Used for Mixture IRT Model

```
# 2 student-level class (G) with prior
# J: the number of students
# I: the number of items
# g: group membership at the student-level
# sigma: the SD of ability
# beta: item difficulty
# eta: ability
# mutg: the mean of ability
# pi: the probability of mixture

model
{
for (j in 1:J) {
    for (i in 1:I) {
       r[j,i]<-resp[j,i]
}}

# Mixture Rasch Model: 2-Group Solution
for (j in 1:J) {
    for (i in 1:I) {
     logit(p[j,i]) <- sigma[g[j]]* eta[j] - b[g[j],i]
     r[j,i]~dbern(p[j,i])
}}

# Ability
for (j in 1:J) {
    eta[j] ~ dnorm(mutg[g[j]], sigt[g[j]])
}
mutg[1] <- 0
mutg[2] ~ dnorm(mut,1)
mut ~ dnorm(0,1)
sigt[1] <- 1
sigt[2] <- 1

# SD of Ability
for (i in 1:G2) {
    sigma[i] ~dnorm(0,1) I(0,)
```

```
}
# Group Membership
for (j in 1:J) {
     g[j] ~ dcat(pi[1:G2])
}
for (g in 1:G2) {
    pi[g] <- delta[g]/sum(delta[1:G2])
    delta[g] ~ dgamma(alpha[g],1)
}
# Item Difficulty
for (g in 1:G2) {
    for (i in 1:I) {
    beta[g,i]~dnorm(0,1)
}}
# Log-Likelihood
for (j in 1:J) {
   for (i in 1:I) {
    l[j,i]<-log(p[j,i])*r[j,i]+log(1-p[j,i])*(1-r[j,i])
 }}
 loglik <-sum(l[1:J,1:I])
 AIC <- -2*(loglik - np)
 BIC <- -2*loglik + np*log(N)
}

list(J=8000, I=40,  G2=2, np=84,
alpha=c(1,1),
resp=structure(.Data=c(
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,0.0,1.0,0.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,0.0,1.0,1.0,0.0,
...
0.0,1.0,1.0,1.0,0.0,
0.0,0.0,1.0,1.0,0.0,
1.0,1.0,0.0,0.0,0.0,
1.0,0.0,0.0,0.0,0.0),.Dim = c(8000,40)))
```

Appendix D

WinBUGS Code Used for MMixIRTM: Prior

```
# 2 student-level class (G) with prior
# 2 school-level class (K)with prior
# J: the number of students
# I: the number of items
# T: the number of schools
# g: group membership at the student-level
# gg: group membership at the school-level
# a: the SD of ability
# b: item difficulty
# eta: ability
# mutg: the mean of ability
model
{
for (j in 1:J) {
    for (i in 1:I) {
     r[j,i] <- resp[j,i]
}}

# G=2

for (j in 1:J) {
    for (i in 1:I) {
     logit(p[j,i]) <-  a[g[j], gg[group[j]]] *eta[j]
     - b[i,g[j],gg[group[j]]]
     r[j,i]~dbern(p[j,i])
}}

# Ability
for (j in 1:J) {
        eta[j]~dnorm(mutg[g[j],gg[group[j]]], sigt[g[j],gg[group[j]]])
}
mutg[1,1] <- 0
mutg[2,1] ~ dnorm(0,1)
mutg[1,2] ~ dnorm(0,1)
mutg[2,2] ~ dnorm(0,1)
sigt[1,1] <- 1
sigt[2,1] <- 1
```

```
sigt[1,2] <- 1
sigt[2,2] <- 1


# SD of Ability
for (g in 1:G2) {
        for (k in 1:K2){
        a[g, k] ~ dnorm(0,1) I(0,)
}}

# Student Level
for (j in 1:N) {
     g[j] ~ dcat(pi[g[group[j]],1:G2])
}

for (k in 1:K2) {
   for (g in 1:G2) {
       pi[k,g] <- delta[k,g] /sum(delta[k,])
       delta[k,g] ~ dgamma(alpha[g],1)
    }
}

# School Level
for (t in 1:T){
    gg[t] ~ dcat(pi1[1:K2])
}
for (k in 1:K2) {
    pi1[k] <- delta1[k]/sum(delta1[1:K2])
     delta1[k] ~ dgamma(alpha1[k],1)
}

# Item Difficulty
for (i in 1:T) {
    for (g in 1:G2) {
        for (k in 1:K2){
        b[i,g,k]~dnorm(0,1)
}}}

# Log-Likelihood
for (j in 1:J) {
   for (i in 1:I) {
    l[j,i]<-log(p[j,i])*r[j,i]+log(1-p[j,i])*(1-r[j,i])
 }}
 loglik <-sum(l[1:J,1:I])
 AIC <- -2*(loglik - np)
```

```
 BIC <- -2*loglik + np*log(N)
}

# Initial Value of School-Level Group Membership
list(gg=c(1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,
1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,
1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,
...
1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,
1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2))

list(J=8000, I=40, T=320, G2=2, K2=2, np=169,
alpha=c(1,1), alpha1=c(1,1),
group=c(1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1
, 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1
, 1   , 1   , 1   , 1   , 2   , 2   , 2   , 2   , 2   , 2   , 2
, 2   , 2   , 2   , 2   , 2   , 2   , 2   , 2   , 2   , 2   , 2
, 2   , 2   , 2   , 2   , 2   , 2   , 2   , 3   , 3   , 3   , 3
...),
resp=structure(.Data=c(
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,0.0,1.0,0.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,0.0,1.0,1.0,0.0,
...
0.0,1.0,1.0,1.0,0.0,
0.0,0.0,1.0,1.0,0.0,
1.0,1.0,0.0,0.0,0.0,
1.0,0.0,0.0,0.0,0.0),.Dim = c(8000,40)))
```

Appendix E

WinBUGS Code Used for MMixIRTM: Multinomial Logistic Regression Model

```
# 2 student-level class (G) with multinomial logistic regression model
# 2 school-level class (K)with multinomial logistic regression model
# J: the number of students
# I: the number of items
# T: the number of schools
# g: group membership at the student-level
# gg: group membership at the school-level
# a: the SD of ability
# b: item difficulty
# eta: ability
# mutg: the mean of ability
# beta's: coefficients of multinomial logistic regression model
model
{
for (j in 1:J) {
    for (i in 1:I) {
     r[j,i] <- resp[j,i]
}}

# G=2
for (j in 1:J) {
    for (i in 1:I) {
     logit(p[j,i]) <-  a[g[j], gg[group[j]]] *eta[j]
     - b[i,g[j],gg[group[j]]]
     r[j,i]~dbern(p[j,i])
}}

# Ability for (j in 1:J) {
        eta[j]~dnorm(mutg[g[j],gg[group[j]]], sigt[g[j],gg[group[j]]])
}
mutg[1,1] <- 0
mutg[2,1] ~ dnorm(0,1)
mutg[1,2] ~ dnorm(0,1)
mutg[2,2] ~ dnorm(0,1)
sigt[1,1] <- 1
```

```
sigt[2,1] <- 1
sigt[1,2] <- 1
sigt[2,2] <- 1

# SD of Ability
for (g in 1:G2) {
        for (k in 1:K2){
        a[g, k] ~ dnorm(0,1) I(0,)
}}

# Student Level
for (j in 1:N) {
     g[j] ~ dcat(pi[gg[group[j]],j,1:G2])
}
for (k in 1:K2) {
  for (j in 1:J) {
    for (g in 1:G2) {
        pi[k,j,g] <- phi[k,j,g] /sum(phi[k,j,])
          log(phi[k,j,g]) <- beta20[k,g]+beta21[g]*gender[j]
          }
    }
}
for (k in 1:K2) {
   for (g in 1:G2){
     beta20[k,g] ~ dnorm(0,1)
    }
}
 for (g in 1:G2){
     beta21[g] ~ dnorm(0,1)
}
# Covariate Identification
beta20[1,1] <- 0
beta20[2,1] <- 0
beta21[1] <- 0
# School Level
for (t in 1:T){
    ggmem2[t] ~ dcat(pi21[t,1:K2])
}
for (t in 1:T){
    for (k in 1:K2) {
         log(phi21[t,k]) <- beta200[k]+ beta211[k]*cito[t]
         pi21[t,k] <- phi21[t,k]/sum(phi21[t,1:K2])
    }
}
for (k in 1:K2){
```

```
     beta200[k] ~ dnorm(0,1)
    beta211[k] ~ dnorm(0,1)
}
# Covariate Identification
beta200[1] <- 0
beta211[1] <- 0
# Item Difficulty
for (i in 1:I) {
    for (g in 1:G2) {
        for (k in 1:K2){
        b2[i,g,k]~dnorm(0,1)
}}}
# Log-Liklihood
for (j in 1:J) {
   for (i in 1:I) {
    l[j,i]<-log(p[j,i])*r[j,i]+log(1-p[j,i])*(1-r[j,i])
 }}
 loglik <-sum(l2[1:J,1:I])
 AIC <- -2*(loglik - np)
 BIC <- -2*loglik + np*log(N)
}

# Initial Value of School-Level Group Membership
list(gg=c(1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,
1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,
1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,
...
1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,
1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2))

list(J=8000, I=40, T=320, G2=2, K2=2, np=169,
group=c(1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1
, 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1   , 1
, 1   , 1   , 1   , 1   , 2   , 2   , 2   , 2   , 2   , 2   , 2
, 2   , 2   , 2   , 2   , 2   , 2   , 2   , 2   , 2   , 2   , 2
, 2   , 2   , 2   , 2   , 2   , 2   , 2   , 3   , 3   , 3   , 3
...),
resp=structure(.Data=c(
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,0.0,1.0,0.0,
1.0,1.0,1.0,1.0,1.0,
1.0,1.0,1.0,1.0,1.0,
1.0,0.0,1.0,1.0,0.0,
```

```
...
0.0,1.0,1.0,1.0,0.0,
0.0,0.0,1.0,1.0,0.0,
1.0,1.0,0.0,0.0,0.0,
1.0,0.0,0.0,0.0,0.0),.Dim = c(8000,40)))
```