A MIXED MEMBERSHIP RASCH MODEL

by

GUOGUO ZHENG

(Under the Direction of Allan Cohen and Hye-Jeong Choi)

ABSTRACT

Mixture IRT models have been applied to investigate the latent groups that exist in the respondent population and how the same set of test items function differently for different latent groups. However, they assume that a respondent remains in the same latent group across test items, which can be unreasonable in certain scenarios. In this dissertation, a mixed membership Rasch model (MMR) is developed to help overcome this limitation in mixture IRT models. The MMR is built by integrating the Rasch model into the framework of mixed membership models which are considered as a soft clustering technique. In the MMR, a respondent belongs to all the latent groups but with different probabilities at the test level. At the item level, a respondent belongs to only one of the latent groups in each test item and the latent group to which he or she belongs can be different across items. For a response to an item, the probability of a correct answer is parameterized using the Rasch model and the item difficulties in the Rasch model are assumed to vary with latent groups. The MMR is estimated using a Metropolis-within-Gibbs algorithm. This dissertation includes three simulation studies. In Study I, parameter recovery of the MMR is investigated given different test conditions and different priors used in the Metropolis-within-Gibbs algorithm, when the item difficulties across latent groups are known. The design and the purpose of Study II are similar to those in Study I except that in Study II,

item difficulties across latent groups are unknown and thus also need to be estimated. In order to run the MMR, the number of latent groups has to be specified even though it is typically unknown. Selecting the best fitting model from among candidate models is an important part of modeling with an MMR. Therefore, in Study III, the performance of several widely applied information criteria is examined in different test conditions in term of their accuracy in selecting the best fitting MMR.

INDEX WORDS: Latent groups, Rasch model, Mixed membership models, Metropolis-within-Gibbs algorithm, Parameter recovery, Model selection

A MIXED MEMBERSHIP RASCH MODEL

by

GUOGUO ZHENG

BA, Bohai University, China, 2012

MA, University of Georgia, 2014

MS, University of Georgia, 2019

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

2021

A MIXED MEMBERSHIP RASCH MODEL

by

GUOGUO ZHENG

Major Professor:	Allan S. Cohen
	Hye-Jeong Choi
Committee:	Nicole Lazar
	Seock-Ho Kim
	Shiyu Wang

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2021

Dedicated to my parents.

They love and support me unconditionally on my long journal of seeking who I am.

ACKNOWLEDGEMENTS

I would like to thank my advisors, Dr. Allan Cohen and Dr. Hye-Jeong Choi. They gave me generous guidance and supports during the entire production of this dissertation. They encouraged me to keep working hard and showed me different perspectives when things did not go smoothly.

I would like to thank my committee, Dr. Nicole Lazar, Dr. Seock-Ho Kim and Dr. Shiyu Wang. They found time in their busy schedules to talk with me whenever I reached out with questions. They also challenged my study designs, results and writing which pushed me to move forward.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER

1 INTRODUCTION 1

2 LITERATURE REVIEW 7

2.1 Rasch Models and Mixture Rasch Models 7

2.2 Mixed Membership Models 12

3 A MIXED MEMBERSHIP RASCH MODEL 17

4 STUDY I 23

4.1 Purpose and Study Design 23

4.2 Data Simulation 24

4.3 Evaluation Statistics for Parameter Recovery 26

4.2 Results 30

5 STUDY II 34

5.1 Purpose and Study Design 34

5.2 Evaluation Statistics for Parameter Recovery 35

5.2 Results 36

6 STUDY III 39

6.1 Purpose and study design....39

6.2 Results....42

7 SUMMARY AND DISCUSSION....44

REFERENCES....49

APPENDICES

A R code for Study I....56

B R code for Study II....59

C R code for Study III....63

LIST OF TABLES

Page

Table 1: Generating item difficulty parameters for a 6-item test.....24

Table 2: Generating item difficulty parameters for a15-item test.....24

Table 3: Parameter recovery in Study I evaluated by average PCR, average RMSE, average bias and average correlation.....32

Table 4: Parameter recovery in study II evaluated by average PCR, average RMSE, average bias, and average correlation.....38

Table 5: Percent of replications in which an information criterion picked out a model as the optimal model.....43

LIST OF FIGURES

Page

Figure 1: The probability density of each dimension of $Dirichlet(0.25,\ 0.25)$.....25

Figure 2: Relationships between posterior estimates and the generating parameters.....29

Figure 3: The distribution of $\widehat{\pi_{i1}}$ and π_{i1} across respondents in one of the replications for sample size *N* = 1000 in Study I.....31

Figure 4: The distribution of $\widehat{\pi_{i1}}$ and π_{i1} across respondents in one of the replications for sample size *N* = 1000 in Study II 36

CHAPTER 1

INTRODCUTION

Item Response Theory (IRT) refers to a family of statistical models designed to measure a continuous latent trait based on observed binary responses to tests, surveys or other types of psychological or educational measures (Baker & Kim, 2004). IRT assumes that the probability of a correct response or the probability of endorsing a target option of an item is a function of item characteristics and the status of the respondent on the latent trait. Item characteristics in the common IRT models include item difficulty, item discrimination and pseudo-guessing. In educational studies, the latent trait is usually defined as the ability in a certain subject, e.g., mathematical computation ability (Wu & Adams, 2006). Based on how many parameters of item characteristics are included, IRT models are categorized as 1-parameter, 2-parameter and 3-parameter models with the Rasch model as a special case of the 1-parameter model.

Another basic assumption of IRT is that a set of items have the same item characteristics for the entire respondent population (Lord & Novick, 1968). This assumption is violated when the respondent population is heterogeneous, such as with respect to the problem-solving strategies used in answering questions (e.g., Mislevy & Verhelst, 1990; Rost, 1990). For example, in a visual spatial task that shows respondents several three-dimensional objects and asks them to determine which of these objects is the same as the target object, respondents may solve the problem by rotating the target object mentally or by detecting matching features between the options and the target object (French, 1965; Lohman, 1979). The same task can be

relatively harder to solve for one strategy than the other depending on the characteristics of the objects (Kyllonen et al., 1984).

The respondent population can also be heterogeneous because of their different test-taking behaviors. One example is from low-stakes tests. Performance in a low-stakes test does not usually have significant consequences for the respondents (Wise & Kong, 2005). The Program for International Student Assessment (PISA), for example, is a low-stakes assessment because its purpose is to evaluate the quality of school systems rather than to make decisions on the students. Even in a low-stakes test, some respondents still actively seek to answer the questions. Wise and Kong describe this kind of test-taking behavior as solution behavior (SB). Other respondents, however, may not necessarily be motivated to answer the questions with high effort. Instead, they may respond so quickly that it is reasonable to assume they did not even fully read and consider the questions. This is called rapid-guessing behavior (RGB). For a given ability level, the same set of questions may appear to be harder for respondents who display RGB as compared with respondents who display SB (Schlosser, Neeman & Attali, 2019).

In the above examples of visual spatial tasks and low-stakes tests, respondents who apply the same problem-solving strategy or who display the same test-taking behavior might be classified into one group. The item characteristics of the same set of test questions may differ across groups. Assuming the same set of item characteristics for the entire population, however, may overestimate or underestimate the item characteristics and ability and even reduce the validity of the tests, since a response is not only a function of the respondent' ability level but also of which latent group to which the respondent belongs (Bolt, Cohen & Wollack, 2002; Embretson, 2007; Oshima, 1994).

When such groups are latent, that is, not observed, mixture IRT models can be used to account for the latent heterogeneity in respondent populations. These models estimate ability and item characteristics for each latent group and also estimate to which latent group each respondent belongs (Rost, 1990). Embretson (2004) and Mislevy and Verhelst (1990) employed mixture IRT models to measure respondent's spatial ability given their problem-solving strategies in visual spatial tasks. Liu et al. (2018) developed a multilevel mixture IRT model for both the process data generated during problem solving and final responses to investigate students' strategy use in a computer-based finding-the-quickest-route task in the Program for International Student Assessment (PISA). Meyer (2010) and Swanson (2015) used mixture IRT and its variants to detect students who randomly guessed answers rather than responded to test questions seriously in low-stakes tests.

Except for being applied to studying problem-solving strategies and test-taking behaviors, mixture IRT models and their extensions have also been applied to studying the effect of speededness (Bolt, Cohen & Wollack, 2002), learning motivation (Johns & Woolf, 2006), food security status (Maia et al., 2020), differential item functioning (Cohen & Bolt, 2005), etc.

Mixture IRT models, following the framework of finite mixture models, assume that each respondent belongs to one and only one of the latent groups across the entire test. They also assume that the probabilities of belonging to each latent group are the same across respondents in the population. When mixture IRT models are applied to studying problem-solving strategies and a latent group represents a strategy, such assumptions may not always be true. There is evidence to suggest that respondents switch strategies across problems and there is individual difference in strategy choice. Young elementary school children, for example, switched strategies on simple arithmetic and spelling problems (Siegler, 1987; Rittle-Johnson & Siegler, 1999), and the

differences in their strategy choice in simple arithmetic problems can be largely explained by the differences in their arithmetic ability (Siegler, 1987). Teenagers were observed to switch strategies at different steps in solving the problem of finding the quickest route in a computer-based PISA task (Liu et al., 2018). In the context of spatial tasks, respondents may switch between strategies across questions, either as an outcome of learning after they answer more questions and explore different strategies (Lohman, 1979), or as motivated by the different characteristics and presentation forms of the questions (Kyllonen, Lohman & Snow, 1984). And the strategy that a respondent tends to employ to effectively solve a spatial task may be a function of the respondent's verbal abilities (Salomon, 1974).

When mixture IRT models are applied to studying test-taking behaviors and a latent group represents a test-taking behavior, the above assumptions in mixture IRT models may not be appropriate, either. For example, factors such as ability, cognitive resources and motivation can predict respondents' general tendency to engage in RGB in low-stakes tests, but engaging in RGB may also be influenced by the characteristics of specific questions, such as the surface features of the questions (Wise et al., 2009). As a result, when the characteristics of the questions change, a respondent may not consistently demonstrate RGB but rather may switch between RGB and SB over the course of the test.

Ignoring respondents' possible multiple membership in latent groups and switching between latent groups across test items might potentially reduce the validity of a test. Mislevy and Verhelst (1990) briefly pointed out this limitation of mixture IRT in a paper on using a mixture IRT model to study problem-solving strategies in visual spatial tasks. However, this issue was not further investigated in subsequent research on mixture IRT models. Therefore, developing a modified IRT model that could account for the possibility that a respondent belongs

to multiple latent groups and switches between latent groups across test items and that also allows for individual differences in the tendency to belong to different latent groups can be useful, especially when the model aims to study respondents' problem-solving strategies or test-taking behaviors while measuring their abilities.

Mixed membership models are a soft clustering technique that allows an individual to belong to multiple latent groups (Erosheva, 2002). In this dissertation, I introduce a mixed membership Rasch (MMR) model developed by integrating a standard Rasch model into the framework of mixed membership models. The literature noted above suggests that, when respondents belong to multiple latent groups and switch between latent groups across test items, it may be a function of the cognitive and noncognitive factors of the respondents or the specific characteristics of the test items. The purpose of the MMR, however, is to account for a simpler scenario of multiple membership and switching behaviors of test respondents using their correct or incorrect responses to test questions. Specifically, this MMR assumes that a respondent belongs to multiple latent groups with different probabilities at test level. These probabilities remain the same for a given respondent across test items but vary over respondents. The model also assumes that in a test item, a respondent belongs to one of the latent groups, but allows a respondent to belong to different latent groups on different test items. This is typically interpreted as the respondent's switching behaviors in the literature of mixed membership models (Erosheva, 2002). Given a latent group, the probability of a correct response to a test item is parameterized using the Rasch model. The item difficulties in the Rasch model are assumed to vary over latent groups.

This dissertation includes seven chapters. In chapter 2, I review mixture IRT models, mixed membership models and their identifiability and scaling issues. In chapter 3, I introduce

the MMR and explain the generative assumptions and the interpretations of this model. I also present the Monte Carlo Markov Chain (MCMC) algorithm used to estimate the model. In order to show that the MMR is a useful model, it is important to understand how well the model can be estimated and what factors would affect the estimation. In chapters 4 and 5, I present the results of simulation studies that investigate how well the parameters in the MMR can be recovered under various simulation conditions. Another practical question to ask before an MMR model is specified is how many latent groups there are in the respondents. This number is usually unknown. Chapter 6, therefore, investigates which model selection indices perform well in selecting the correct number of latent groups for the MMR models using simulated data. In chapter 7, I summarize the findings of chapters 4-6 and discuss the scaling issue in the MMR and the challenges of comparing current results with those in previous related studies on IRT. I also discuss what might be investigated in future studies for the MMR.

CHPATER 2

LITERATURE REVIEW

2.1 Rasch Models and Mixture Rasch Models

In IRT, the probability of a correct response to a dichotomous item is a function of item characteristics and the status of the respondent on the latent trait. This latent trait is commonly referred to as ability in education studies. In the Rasch model, item difficulty is assumed to vary with test items, item discrimination is fixed as 1 across items and the probability of correctly guessing on an item when ability level is extremely low is assumed to be zero (Baker & Kim, 2004). Let θ_i denote the ability of respondent *i*, b_j denote the difficulty of item *j* and X_{ij} denote a binary response of respondent *i* to question *j*. In the Rasch model, the probability of a correct response to item *j* is:

$$P(X_{ij} = 1| \theta_i, b_j) = \frac{1}{1+ e^{-(\theta_i - b_j)}}$$

and X_{ij} follows a Bernoulli distribution with success probability $P(X_{ij} = 1| \theta_i, b_j)$. Suppose there are *N* respondents and *J* items. Given $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_i, \ldots, \theta_N)$ and $\boldsymbol{b} = (b_1, \ldots, b_j, \ldots, b_J)$, responses to all the items from all the respondents are assumed to be independent. In this case, the likelihood of the responses across respondents and across items given ability and item difficulty parameters is the product of the probability of a response given the relevant ability and item difficulty parameters, $P(\mathbf{X}| \boldsymbol{\theta}, \boldsymbol{b}) = \prod_i \prod_j P\left(X_{ij} \middle| \theta_i, b_j\right)$.

Mixture Rasch models follow the framework of finite mixture models. These models assume that there are *G* latent groups in the respondent population and the item difficulties in the

Rasch model vary across latent groups (Bolt, Cohen & Wollack, 2002). The mixture Rasch model can be specified as follows:

Suppose there are *G* latent groups, let b_{gj} denote the item difficulty of item *j* in latent group *g*, θ_{gi} denote respondent *i*' s ability given membership in latent group *g*, and Z_i denote the latent group to which respondent *i* belongs. Z_i can take on integer values that range from 1 to *G*. b_{gj} and θ_{gi} are latent group specific and thus are the component parameters in a mixture Rasch model. In a mixture Rasch model, the probability of a correct response to item *j* given latent group *g* and the corresponding component parameters is the following:

$$P(X_{ij} = 1 | Z_{ig} = g, \theta_{gi}, b_{gj}) = \frac{1}{1+ e^{-(\theta_{gi} - b_{gj})}}$$

It is noted that even though the ability parameter has a latent group index, only one ability is estimated for each respondent since mixture Rasch models assume that a respondent belongs to only one of the latent groups. One of the possible reasons of using a latent group index on ability parameters is that in some scenarios, the distribution of ability is assumed to vary with latent groups and having a latent group index makes it straightforward to assign a different prior distribution for the abilities in different latent groups. Another possible reason is to indicate that the estimation of a respondent's ability is a function of the latent group to which he or she belongs and the item difficulty parameters in that latent group. Such information is needed to develop an estimation algorithm for a mixture Rasch model. Given $\boldsymbol{\theta}$, $\boldsymbol{b}$ and $\boldsymbol{Z}$ = (Z_1,…, Z_N), responses to all the items from all the respondents are assumed to be independent with the likelihood of all the responses written as $P(\mathbf{X}| \boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z}) = \prod_i \prod_j P\left(X_{ij} \middle| \theta_{gi}, b_{gj}, Z_{ig} = \mathrm{g}\right)$.

Scaling in the Rasch and Mixture Rasch Models. In the Rasch model, the scale of item difficulty and ability parameters is undetermined. What this means is that a linear transformation of a set of item difficulty and ability parameters in the Rasch model can return the same

probability of a correct response given the original set of the parameters (Kolen & Brennan, 2014). Let θ_i^* and b_j^* denote the rescaled θ_i and b_j after a linear transformation. Equation (1) shows that both (θ_i, b_j) and (θ_i^*, b_j^*) return the same probability of a correct response since $P(X_{ij} = 1|\theta_i^*, b_j^*)$ equals $P(X_{ij} = 1|\ \theta_i, b_j)$:

$$\theta_i^* = \theta_i + B \qquad (1)$$

$$b_j^* = b_j + B$$

$$P(X_{ij} = 1|\theta_i^*, b_j^*) = \frac{1}{1+ e^{-(\theta_i^* - b_j^*)}}$$

$$= \frac{1}{1+ e^{-(\theta_i + B\ - b_j - B)}}$$

$$= \frac{1}{1+ e^{-(\theta_i - b_j)}}$$

$$= P(X_{ij} = 1|\ \theta_i, b_j)$$

Therefore, the scale of the item difficulty and ability parameters in the Rasch model is not unique.

In order to provide a scale for the item difficulty and ability parameters and also to ensure that the parameters across latent groups are on the same scale and thus are comparable, researchers typically use the constraint $\sum_j b_{gj} = 0$ for item difficulties within each latent group in mixture Rasch models (Bolt, Cohen & Wollack, 2002; Meyer, 2010; Mislevy & Verhelst, 1990).

Label switching in Bayesian estimation of the mixture Rasch models. Mixture Rasch models as well as finite mixture models in general have the issue of lack of identifiability associated with the permutations of latent group indices. That is, there is more than one way to label the latent groups and the different ways of labeling would return the same likelihood or posterior distributions in certain cases. As a result, label switching, which means latent groups switch between indices, may occur in the estimation algorithms of these models. The following

equations illustrate this lack of identifiability in mixture Rasch models estimated using Bayesian methods.

In mixture Rasch models, the likelihood of all the respondents' responses to all the items is $P(\mathbf{X}|\,\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z}) = \prod_i \prod_j P(X_{ij}|\,\theta_{gi}, b_{gj}, Z_{ig} = g)$. Let γ denote a permutation of latent group indices. Equation (2) shows that this likelihood is the same before and after a permutation of the latent group indices, that is, $P(\mathbf{X}|\,\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z}) = \mathrm{P}(\mathbf{X}|\,\gamma(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z}))$. To understand this idea intuitively, suppose there are two latent groups. Whether the first latent group is being called group 1 or the second latent group is being called group 1 does not affect the likelihood of the responses as long as such naming is consistent across latent-group-specific parameters. Therefore, the likelihood of a mixture Rasch model is only identifiable up to a permutation of the latent group indices.

$$
\begin{aligned}
P(\mathbf{X}|\,\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z}) &= \prod_i \prod_j P(X_{ij}|\,\theta_{gi}, b_{gj}, Z_{ig} = g) \qquad (2)\\
&= \prod_i \prod_j P(X_{ij}|\,\theta_{\gamma(g)i}, b_{\gamma(g)j}, Z_i = \gamma(g)))\\
&= \mathrm{P}(\mathbf{X}|\,\gamma(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z}))
\end{aligned}
$$

In a mixture Rasch model, when priors are exchangeable over latent group indices, the posteriors are also only identifiable up to a permutation of the latent group indices. In statistics, exchangeability means the joint distribution of a set of variables $\mathrm{P}(Y_1, Y_2, \ldots, Y_s)$ does not change after a permutation of the indices $\mathrm{P}(Y_{\omega(1)}, Y_{\omega(2)}, \ldots, Y_{\omega(s)})$ where ω denotes a permutation. With respect to the mixture Rasch models, the joint prior distribution of all the parameters can be written as the product of the prior distribution of each parameter since the prior distributions are assumed to be independent:

$$
\begin{aligned}
\mathrm{P}(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z}) &= \mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\boldsymbol{b})\mathrm{P}(\boldsymbol{Z})\\
&= \prod_{i=1}^{N} \prod_{g=1}^{G} P(\theta_{gi})^{I(Z_i=g)} \prod_{j=1}^{J} \prod_{g=1}^{G} P(b_{gj}) \prod_{i=1}^{N} \prod_{g=1}^{G} P(Z_i = g)^{I(Z_i=g)}
\end{aligned}
$$

With a permutation of latent group indices, this prior can be written as:

$$\mathrm{P}(\gamma(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z})) = \mathrm{P}(\gamma(\boldsymbol{\theta}))\mathrm{P}(\gamma(\mathbf{b}))\mathrm{P}(\gamma(\mathbf{Z}))$$

$$= \prod_{i=1}^{N}\prod_{g=1}^{G} P(\theta_{\gamma(g)i})^{I(Z_i=\gamma(g))} \prod_{j=1}^{J}\prod_{g=1}^{G} P(b_{\gamma(g)j}) \times$$

$$\prod_{i=1}^{N}\prod_{g=1}^{G} P(Z_i = \gamma(g))^{I(Z_i=\gamma(g))}$$

If this prior distribution does not change after a permutation of the latent group indices, that is, $\mathrm{P}(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z}) = \mathrm{P}(\gamma(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z}))$, we say this prior is exchangeable. An example of an exchangeable prior in a mixture Rasch model is when the priors on b_{gj}, θ_{gi} and Z_i are the same regardless of g.

Equation (3) shows the posterior distribution is the same before and after the permutation of latent group indices with an exchangeable prior, that is, $\mathrm{P}(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z}|\boldsymbol{X}) = P(\gamma(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z})|\boldsymbol{X})$.

$$P(\gamma(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z})|\boldsymbol{X}) \qquad (3)$$

$$= \frac{\mathrm{P}(\mathbf{X}|\,\gamma(\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{Z}))\,\mathrm{P}(\gamma(\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{Z}))}{P(\boldsymbol{X})}$$

$$= \frac{\mathrm{P}(\mathbf{X}|\,\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{Z})\,\mathrm{P}(\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{Z})}{P(\boldsymbol{X})}$$

$$= \mathrm{P}(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{Z}|\boldsymbol{X})$$

Therefore, the posterior in a mixture Rasch model is only identifiable up to a permutation of the latent group indices with an exchangeable prior and the latent groups may switch between indices within or across estimation algorithms of a mixture Rasch model even for the same data set. When the mixture Rasch models are estimated using MCMC, multiple modes may display among the samples obtained in a single MCMC chain or in multiple MCMC chains.

When label switching occurs, it is inappropriate to make inferences about each parameter using the posterior samples directly. This is because the posterior distributions of the model parameters, such as $\boldsymbol{b}$ in the mixture Rasch models, are not distinguishable across latent groups

(Gelman et al., 2013). Label switching is not only an issue for mixture Rasch models but for finite mixture models in general.

In order to handle the label switching issue in mixture Rasch models, some researchers have placed ordinal constraints on the parameters or used non-exchangeable priors. These methods have the potential to cause the invariance property of the posterior distributions under the permutation of indices to be violated (Bolt, Cohen & Wollack, 2002; Huang, 2016; Meyer, 2010; Sen, Cohen & Kim, 2016). Choosing such constraints or priors in these studies has usually relied on reasonable assumptions about the data or study design. Post-processing methods have also been used to remove label switching across MCMC chains for mixture IRT models when it is observed (Cho, Cohen & Kim, 2013; Choi & Wilson, 2014; Finch, 2012). In post-processing, researchers usually examine whether obvious jumps are evident in the trace plots across multiple MCMC chains. Such jumps may indicate the occurrence of label switching. When they are observed, posterior samples might be manually relabeled so that the indices are consistent across multiple MCMC chains.

2.2 Mixed Membership Models

Mixed membership modeling is a general framework that has incorporated previous statistical models reflecting the idea that individuals may belong to multiple latent groups (Erosheva, 2002; Erosheva et al., 2004; Galyardt, 2012). It differs from finite mixture models in that it does not assume an individual belongs to one and only one of the latent groups in the data. The mixed membership models are not used to measure abilities and do not include parameters for abilities nor for item characteristics as IRT does. Instead, they have been applied in a wide range of contexts to investigate the distributions of observed variables under different latent groups and individuals' multiple memberships in the latent groups. Erosheva et al. (2004), for example, used

a mixed membership model adjusted for text data to study what research areas were mentioned in the biology articles published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS) from 1997 to 2001, how likely different words would appear in an article given an area and how much an article expressed each of the areas. In the setting of education, Galyardt (2012) combined a mixed membership model with a response time model to study what strategies young children used to solve least common multiples problems and how much each child used each of the strategies in the test.

Mixed membership models and their variants have also been applied to studying the multiple disease profiles in patients (Woodbury, Clive & Garson, 1978), multiple genetic heritages in birds (Pritchard, Stephens & Donnelly, 2000) and multiple communities a monk at a monastery has social interactions with (Airoldi et al., 2008).

There are four levels of assumptions in mixed membership models (Erosheva, 2002; Galyardt, 2012). These four levels of assumptions are illustrated below and the contexts of Erosheva et al. (2004) and Galyardt (2012) are used as examples to explain the notations mentioned.

Population level. It is assumed that there are G latent groups in the population. The interpretations of the latent groups depend on data and study design. For example, in Erosheva et al. (2004), latent groups are the research areas mentioned in the articles published in PNAS. In Galyardt (2012), latent groups are the strategies that young children used to solve least common multiples problems.

The distribution of an observed variable varies with latent groups. Let variables be indexed by $j = 1, \ldots, J$ and latent groups by $g = 1, \ldots, G$. Given latent group g, the distribution of variable j is denoted by $P_g(X_j)$. For Erosheva et al. (2004), variables are the words that appear in

an article and $P_g(X_j)$ denotes a multinomial distribution of words with a single trial given research area *g*. For Galyardt (2012), variables are children's responses to least common multiple problems and $P_g(X_j)$ denotes a Bernoulli distribution of a right or wrong response to a problem given strategy *g*.

Individual level. Each individual has a membership probability vector $\boldsymbol{\pi_i} = (\pi_{i1}, \ldots, \pi_{ig}, \ldots, \pi_{iG})$, where π_{ig} denotes the probability that individual *i* belongs to latent group *g*. Each element in this vector is nonnegative and falls between 0 and 1, and all the elements in a given vector sum to 1. $\boldsymbol{\pi_i}$ indicates individual *i*'s partial membership in each of the latent groups. For example, suppose we have two latent groups, $G = 2$, and an individual has $\boldsymbol{\pi_i} = (0.30, 0.70)$. This individual belongs to latent group 1 with probability 0.3 and to latent group 2 with probability 0.7. In the context of Erosheva et al. (2004), $\boldsymbol{\pi_i} = (0.30, 0.70)$ would mean that article *i* expresses one of the two areas with probability 0.30 and the other area with probability 0.70. And in the context of Galyardt (2012), it would mean child *i* uses one of the two strategies 30% of the time and the other strategy 70% of the time in the test.

Let Z_{ij} denote the latent group that individual *i* belongs to in variable *j*. Z_{ij} can take on integer values that range from 1 to *G*. $Z_{ij} = g$ is used to denote that individual *i* belongs to latent group *g* in variable *j*. $\boldsymbol{Z_i} = (Z_{i1}, \ldots, Z_{ij}, \ldots, Z_{iJ})$ indicates the behavior that individual *i* switches latent groups across variables. For Erosheva et al. (2004), Z_{ij} would indicate the research area that word *j* in article *i* represents and different words in article *i* can represent different areas. For Galyardt (2012), Z_{ij} would indicate the strategy child *i* uses to solve problem *j* and child *i* may use different strategies to solve different problems in the test.

Given $\boldsymbol{\pi_i}$, the marginal response distribution of individual *i* for variable *j* is

$$P(X_j|\boldsymbol{\pi_i}) = \sum_{g=1}^{G} P\left(X_j \middle| Z_{ij} = g\right) P(Z_{ij} = g|\boldsymbol{\pi_i}) = \sum_{g=1}^{G} \pi_{ig}\, P_g(X_j)$$

Given $\boldsymbol{\pi_i}$, individual *i*'s responses to all the variables are assumed to be independent, $P(\mathbf{X}|\boldsymbol{\pi_i}) = \prod_{j=1}^{J} \sum_{g=1}^{G} \pi_{ig} P_g(X_j)$.

Sampling scheme level. In some cases, each variable is measured repeatedly for each individual. The number of repeated measurements can be different across variables and across individuals. Let R_{ij} denote the number of repeated measurements of variable *j* for individual *i*. An individual's responses at all repeated measurements across variables are assumed to be independent given $\boldsymbol{\pi_i}$, $P(\mathbf{X}|\boldsymbol{\pi_i}) = \prod_{j=1}^{J} \prod_{r=1}^{R_{ij}} \sum_{g=1}^{G} \pi_{ig} P_g(X_{jr})$. In Erosheva et al. (2004), the observed variables are what words appear in an article, and the distribution of words given a research area, $P_g(X_j)$, is assumed to be the same regardless of the location of the words. Suppose there are *R* words in article *i*. The *R* words can be considered as *R* repeated measurements of what word appears in the article. And the number of variables, *J*, can be considered as 1.

Latent variable level. Latent variables in the context of mixed membership models refer to $\boldsymbol{\pi_i}$, where $\boldsymbol{\pi_i}$ can be treated either as fixed but unknown constants or random samples from a certain underlying distribution for the purpose of estimation. When treated as random, $\boldsymbol{\pi_i}$ is usually assumed to follow a Dirichlet distribution in which the components are independent subject to the constraint $\sum_{i=1}^{G} \pi_{ig} = 1$ (Aitchison, 1982). Or $\boldsymbol{\pi_i}$ may be assumed to follow a logistic normal distribution, where covariances between the elements in $\boldsymbol{\pi_i}$ are explicitly indicated in the probability density function. Blei and Lafferty (2007) used the logistic normal distribution, for example, in a mixed membership model adjusted for text data to model correlated topics in text documents.

Label Switching in Mixed Membership Models. Similar to finite mixture models, mixed membership models also have an issue of lack of identifiability. Galyardt (2012) proved that a general mixed membership model can be expressed as a finite mixture model that has a much

larger number of latent groups compared with the equivalent mixed membership model and has constraints on membership probabilities. Further, Galyardt (2012) proved that mixed membership models are only identifiable up to a permutation of the latent group indices when the components of the $\boldsymbol{\pi_i}$ are exchangeable, as is the case when a symmetric Dirichlet prior is placed on $\boldsymbol{\pi_i}$. Therefore, when a mixed membership model is estimated using MCMC, label switching may occur in the posterior samples either within a single MCMC chain or across multiple MCMC chains.

Some researchers who applied mixed membership models in their studies have employed similar methods as those mentioned in mixture IRT studies to handle this label switching issue, such as imposing ordinal constraints on $\boldsymbol{\pi_i}$ or latent-group-specific parameters (Richardson & Green, 1997) and post-processing the posterior samples so that the latent group labels are consistent across estimation algorithms or between the posterior estimates and the generating values in simulation studies (Wang & Erosheva, 2015).

CHAPTER 3

A MIXED MEMBERSHIP RASCH MODEL

Mixture IRT models assume each respondent remains in the same latent group across the entire test. As noted in Chapter 1, such an assumption may not always be reasonable especially when the mixture IRT models are used to study problem-solving strategy use and test-taking behaviors in low-stakes tests. To help tackle this limitation in mixture IRT models, a mixed membership Rasch model (MMR) is developed in this dissertation. It integrates the Rasch model into the mixed membership framework. In the MMR, an individual is a respondent and observed variables are responses to the items in a test. The distribution of a response to an item by a respondent is parameterized using the Rasch model given a latent group. The item difficulty parameters in the mixture Rasch model are assumed to vary with latent groups. The MMR allows a respondent to belong to multiple latent groups with different probabilities at the test level and to belong to different latent groups on different items in a test.

The generative process of the MMR is explained as follows:

1. Assume there are *G* latent groups in the sample of respondents. Again, the interpretations of the latent groups depend on data and study design. For example, if the purpose is to investigate students' test-taking behaviors in a low-stakes test, latent groups may correspond to solution behavior and random guessing behavior even though such interpretation may need to be supported by further evidence, e.g., by cognitive interviews.
2. Each respondent *i* has a membership probability vector $\boldsymbol{\pi_i}$ of length *G*. π_{ig} indicates the probability that respondent *i* belongs to latent group *g*. The elements of $\boldsymbol{\pi_i}$ are nonnegative

and fall within the range 0 and 1 with the sum of all the elements in $\boldsymbol{\pi_i}$ equal to 1. The MMR assumes that $\boldsymbol{\pi_i}$ is drawn from a Dirichlet distribution:

$$\boldsymbol{\pi_i} \sim Dirichlet(\boldsymbol{\alpha}) \quad \boldsymbol{\pi_i} = (\pi_{i1}, \ldots, \pi_{ig}, \ldots, \pi_{iG})$$

Using the test-taking behavior example, suppose latent group 1 represents solution behavior (SB) and latent group 2 represents random guessing behavior (RGB). If respondent *i* has a $\boldsymbol{\pi_i}$ = (0.2, 0.8), this respondent would show solution behavior with probability 0.2 and random guessing behavior with probability 0.8 in the test. The MMR assumes that these probabilities are different for different respondents. It also assumes that these probabilities for respondent *i* remain the same regardless of which question in the test he or she is trying to answer.

3. In item *j*, the latent group to which respondent *i* belongs is denoted by Z_{ij} and is drawn from a multinomial distribution with probabilities $\boldsymbol{\pi_i}$:

$$Z_{ij}|\boldsymbol{\pi_i} \sim Multinomial(1, \boldsymbol{\pi_i})$$

The latent group respondent *i* belongs to may vary across items. Therefore, the Z_{ij}'s indicate respondent *i*'s switching behaviors. In the example of test-taking behaviors, Z_{ij} denotes the type of test-taking behavior respondent *i* shows on item *j* and Z_{ij}'s indicate that the respondent *i* switches between SB and RGB across test questions.

4. Let X_{ij} denote a binary response by respondent *i* to item *j*, θ_i denote respondent *i*'s ability and b_{gj} denote item *j*'s difficulty given latent group *g*. X_{ij} takes on value 1 for a correct response and 0 otherwise. Given latent group *g*, θ_i and b_{gj}, X_{ij} is generated from a Bernoulli distribution with probability P_{ijg}:

$$X_{ij}|Z_{ij} = g, \theta_i, b_{gj} \sim Bernoulli(P_{ijg})$$

$$P_{ijg} = \frac{1}{1+ e^{-(\theta_i - b_{gj})}}$$

With respect to the example of test-taking behaviors, this indicates how likely a respondent answers a question correctly depends on his or her ability level, how difficult the question is and whether he or she answers the question using SB or RGB.

It is noted that in the general mixed membership models, the distribution of an observed variable can change for different variables and latent groups. In the MMR, an observed variable is a binary response to a test item. Its distribution can change for different test items, latent groups and respondents.

In this study, the MMR is estimated using a Metropolis-within-Gibbs algorithm. In a Metropolis-within-Gibbs algorithm, some parameters are sampled using Gibbs sampler and some are sampled using Metropolis algorithm. Both the Gibbs sampler and the Metropolis algorithm are Markov Chain Monte Carlo (MCMC) sampling methods. The MCMC draws a sequence of samples of parameters from approximate distributions as running a Markov chain in which, the stationary distribution is the target posterior distribution. The specific steps of the Metropolis-within-Gibbs algorithm are as follows:

Step 1. For each respondent, $\boldsymbol{\pi_i}$ has the following posterior distribution assuming conditional independence of respondent *i*'s responses across items:

$$\mathrm{P}(\boldsymbol{\pi_i}|\text{rest}) \propto \prod_j \mathrm{P}(Z_{ij}|\boldsymbol{\pi_i}) \times \mathrm{P}(\boldsymbol{\pi_i})$$

$$\propto \prod_g \pi_{ig}^{\sum_j \mathrm{I}(Z_{ij}=g)} \prod_g \pi_{ig}^{\alpha_g}$$

$$= \text{Dirichlet}(\boldsymbol{\alpha}^*)$$

$$\alpha_g^* = \alpha_g + \sum_j \mathrm{I}(Z_{ij} = g)$$

where "rest" is shorthand for data and the rest of the parameters in the MMR, and P($\boldsymbol{\pi_i}$) denotes the prior distribution of $\boldsymbol{\pi_i}$ and is a Dirichlet distribution with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_G)$. The posterior distribution of $\boldsymbol{\pi_i}$, P($\boldsymbol{\pi_i}$|rest), is a Dirichlet distribution with $\boldsymbol{\alpha}^* = (\alpha_1^*, \ldots, \alpha_G^*)$.

At iteration *t*, draw a sample of $\boldsymbol{\pi_i}$ from P($\boldsymbol{\pi_i}$|rest) given the samples of the other parameters obtained at iteration *t-1*.

Step 2. For each respondent and each item, Z_{ij} has the following the posterior distribution:

$$\begin{aligned} \mathrm{P}(Z_{ij}|\text{rest}) &\propto \mathrm{P}(X_{ij}|Z_{ij}, \theta_i, b_{gj}) \times \mathrm{P}(Z_{ij}|\boldsymbol{\pi_i}) \\ &\propto \prod_g \left[{P_{ijg}}^{X_{ij}} (1 - P_{ijg})^{1-X_{ij}} \right]^{\mathrm{I}(Z_{ij}=g)} \prod_g {\pi_{ig}}^{\mathrm{I}(Z_{ij}=g)} \\ &= \prod_g \left[{P_{ijg}}^{X_{ij}} (1 - P_{ijg})^{1-X_{ij}} \pi_{ig} \right]^{\mathrm{I}(Z_{ij}=g)} \end{aligned}$$

where $\mathrm{I}(Z_{ij} = g)$ is an indicator function taking on value 1 if respondent *i* belongs to latent group *g* on item *j* and 0 otherwise. The posterior distribution of Z_{ij}, P(Z_{ij}|rest), is a multinomial distribution with probabilities ${P_{ij1}}^{X_{ij}} (1 - P_{ij1})^{1-X_{ij}} \pi_{i1}, \ldots, {P_{ijG}}^{X_{ij}} (1 - P_{ijG})^{1-X_{ij}} \pi_{iG}$.

At iteration t, draw a sample of Z_{ij} from P(Z_{ij}|rest) given the samples of the other parameters obtained at iteration *t-1*.

Step 3. Assuming conditional independence of respondent *i*'s responses across items, the full conditional distribution of θ_i is as follows:

$$\begin{aligned} \mathrm{P}(\theta_i|\text{rest}) &\propto \prod_j \mathrm{P}(X_{ij}|Z_{ij}, \theta_i, b_{gj}) \times \mathrm{P}(\theta_i) \\ &\propto \prod_j \left[\prod_g {P_{ijg}}^{X_{ij}} (1 - P_{ijg})^{1-X_{ij}} \right]^{\mathrm{I}(Z_{ij}=g)} Normal(\theta_i;\ \eta, \sigma^2) \end{aligned}$$

where $Normal(\theta_i;\ \eta, \sigma^2)$ is the normal density evaluated at θ_i given mean η and variance σ^2. θ_i is updated using a Metropolis step, since the normal distribution is not a conjugate prior. In order to obtain a sample of θ_i at iteration *t*,

- Draw a proposed value θ_i^* from *Normal*(θ_i^{t-1},1) where θ_i^{t-1} is a sample of θ_i obtained at iteration *t-1*.
- Calculate the ratio of the posterior densities given the samples of the other parameters that are obtained at iteration *t-1*:

$$r_{\theta_i^*} = \frac{P(\theta_i^*|rest)}{P(\theta_i^{t-1}|rest)}$$

- Assign $\theta_i^t = \theta_i^*$ with probability min{1, $r_{\theta_i^*}$}, and assign $\theta_i^t = \theta_i^{t-1}$ otherwise.

Step 4. Assuming conditional independence of the responses to item *j* across respondents, the full conditional distribution of b_{gj} is as follows:

$$\mathrm{P}(b_{gj}|\ \text{rest}) \propto \prod_i \mathrm{P}(\ X_{ij}|Z_{ij}, \theta_i, b_{gj}) \times \mathrm{P}(b_{gj})$$

$$\propto \prod_i \left[\prod_g {P_{ijg}}^{X_{ij}}\ (1 -\ P_{ijg})^{1-X_{ij}}\right]^{\mathrm{I}(Z_{ij}=g)}\ Normal(b_{gj};\ \mu, \tau^2)$$

where *Normal*(b_{gj}; μ, τ^2) is the normal density evaluated at b_{gj} with mean μ and variance τ^2. b_{gj} is updated using a Metropolis step since the normal distribution is not a conjugate prior. In order to obtain a sample of b_{gj} at iteration *t*,

- Draw a proposed value b_{gj}^* from *Normal*(b_{gj}^{t-1},1) where b_{gj}^{t-1} is a sample of b_{gj} obtained at iteration *t*-1.
- Calculate the ratio of the posterior densities given the samples of the other parameters obtained at iteration *t-1*:

$$r_{b_{gj}^*} = \frac{P(b_{gj}^*|rest)}{P(b_{bj}^{t-1}|rest)}$$

- Assign $b_{gj}^t = b_{gj}^*$ with probability min{1, $r_{b_{gj}^*}$ }, and assign $b_{bj}^t = b_{bj}^{t-1}$ otherwise.

The scales of the parameters in the Rasch model are undetermined, as described in section 2.1. In order to set up a scale for the parameters and also to ensure that the parameters

across latent groups are on the same scale and thus are comparable, after item difficulties are sampled for all the items in latent group g at each iteration of MCMC, the sampled item difficulties are rescaled so that $\sum_j \widehat{b_{gj}} = 0$. For example, suppose the original item difficulty samples across J items at an iteration sum up to a, $\sum_j \widehat{b_{gj}}^* = a$. After the rescaling $\widehat{b_{gj}}^* - \frac{a}{J}$, the sum of the rescaled samples equals 0, $\sum_j \left(\widehat{b_{gj}}^* - \frac{a}{J}\right) = 0$.

CHAPTER 4

STUDY I

4.1 Purpose and Study Design

Both Study I and Study II are to investigate how well the parameters in a two-latent-group MMR can be recovered under different conditions. In study I, it is assumed that there is extra information available about item difficulty parameters. Compared with a mixture Rasch model, an MMR with the same number of latent groups is a more complicated model and has far more parameters to estimate. The investigation of parameter recovery in Study I, therefore, starts with a scenario in which item difficulty parameters are known. This is analogous to having an item bank, in which the item difficulty parameters are known. Item difficulties, thus, do not need to be estimated.

The parameter recovery of the MMR is investigated under three conditions. The first condition is test length: 6-item, 15-item and 30-item tests were simulated to reflect very small, small and medium test lengths. The generating item difficulty parameters for the two latent groups used in this study, as shown in Table 1 and Table 2, are taken from Li, Cohen, Kim and Cho (2009). In the 30-item test, each item in the 15-item test is repeated once. The sum of these parameters equals zero within each latent group. The second condition is sample size: 300, 500 and 1000 respondents' responses were simulated to reflect small, medium and large sample sizes common in education studies. The third condition is the prior choice for the membership probability parameters $\boldsymbol{\pi_i}$ The effects of two priors were tested. One prior was the same as the generating distribution (i.e., the true prior) and the other prior was a flat prior to reflect a lack of

prior knowledge about the distribution of $\boldsymbol{\pi_i}$. This is further explained under Priors Used in Both Study I and Study II (see below).

Table 1. Generating item difficulty parameters for the 6-item test

Item	Item difficulty	
	Latent Group 1	Latent Group 2
1	-1.50	0.00
2	-1.50	0.00
3	0.00	1.50
4	0.00	1.50
5	1.50	-1.50
6	1.50	-1.50

Table 2. Generating item difficulty parameters for the 15-item test

Item	Item difficulty	
	Latent Group 1	Latent Group 2
1	-2.00	-0.50
2	-1.75	-0.25
3	-1.50	0.00
4	-1.25	0.25
5	-1.00	0.50
6	-0.50	1.00
7	-0.25	1.25
8	0.00	1.50
9	0.25	1.75
10	0.50	2.00
11	1.00	-2.00
12	1.25	-1.75
13	1.50	-1.50
14	1.75	-1.25
15	2.00	-1.00

4.2 Data Simulation

The data of different test lengths and samples sizes were simulated following this procedure:

- Choose a test length *J* which varies by conditions
- Choose a number of respondents *N* which varies by conditions
- Number of latent groups *G*= 2

- For each latent group g = 1, 2, generating item difficulty parameters are shown in Table 1 and Table 2. In the 30-item test, each item in the 15-item test is repeated once.
- For each respondent i = 1, …, N, simulate
  a. $\boldsymbol{\pi_i} \sim Dirichlet(0.25,\ 0.25)$. In $Dirichlet(0.25,\ 0.25)$, about 37% of the data have a π_{i1} smaller than 0.20 and a π_{i2} larger than 0.80, and about 37% of the data have a π_{i1} larger than 0.80 and a π_{i2} smaller than 0.20. Therefore, using this generating distribution for $\boldsymbol{\pi_i}$ simulates the scenario that most of the respondents tend to have a dominant latent group and tend to stay in that latent group across items.

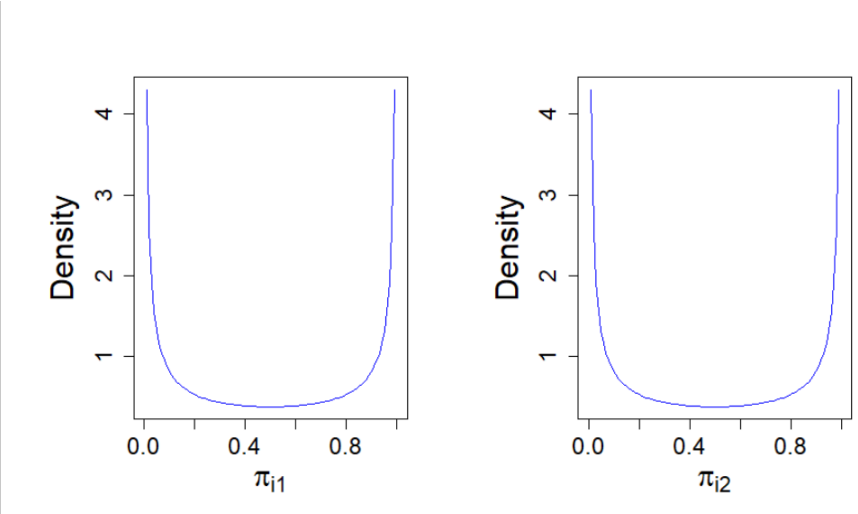

Figure 1. The probability density of each dimension of $Dirichlet(0.25,\ 0.25)$. Most of the data either have a large π_{i1} and a small π_{i2} or a small π_{i1} and a large π_{i2}.

  b. $\theta_i \sim$ N (0, 1)
  c. For each item j = 1, …, J, simulate
     $Z_{ij}|\boldsymbol{\pi_i} \sim Multinomial(\boldsymbol{\pi_i})$

     $X_{ij}|Z_{ij} = g \sim Bernoulli(P_{ijg})$ where $P_{ijg} = \frac{1}{1+e^{-(\theta_i - b_{gj})}}$

The priors used in both Study I and Study II are as follows:

- $\theta_i \sim N(0, 1)$
- $b_{gj} \sim N(0, 1)$
- Two priors of $\boldsymbol{\pi_i}$, P($\boldsymbol{\pi_i}$)

  Prior 1: the true prior which is the same as the generating distribution of $\boldsymbol{\pi_i}$, $Dirichlet(0.25,\ 0.25)$

  Prior 2: a flat prior, $Dirichlet(1,\ 1)$

The simulation of response data was replicated 50 times under each test length and each sample size. And in the same condition, the same set of generating parameters was used across replications. To illustrate the simulated data, assuming a simulated student responds to 15 questions, his or her data simulated based on the above process would look like the following:

$$\theta_1 = -1.32$$

$$\boldsymbol{\pi_1} = (0.56, 0.44)$$

$$\boldsymbol{Z_1} = (1, 1, 2, 1, 2, 1, 2, 1, 2, 2, 1, 1, 2, 1, 1)$$

$$\boldsymbol{Y_1} = (0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0)$$

The total number of conditions under which parameter recovery was examined was therefore 18 (i.e., 3 test lengths × 3 sample sizes × 2 priors of membership probabilities = 18 conditions). After running MCMC for each data set, convergence and label switching were examined.

4.3 Evaluation Statistics for Parameter Recovery

Under each condition, the recoveries of membership probability $\boldsymbol{\pi_i}$ and ability θ_i were evaluated using the average root mean square error (RMSE), average bias and the average correlation between posterior estimates and the true parameters. The recovery of latent group membership

Z_{ij} was evaluated using the average proportion of correct recovery (PCR). For $\boldsymbol{\pi_i}$, because π_{i1} + π_{i2} = 1, the patterns of the estimates of π_{i1} and π_{i2} are reversed. Therefore, the following analyses only focused on π_{i1}.

The mean squared error (MSE) of an estimator is the expected squared difference between an estimate and the true parameter and can be written as the sum of two components, bias^2 and variance. The RMSE is the square root of the MSE. In this study, the RMSE's for $\boldsymbol{\pi_i}$ and θ_i are averaged over respondents respectively. The mathematical definitions of the average RMSE's of $\boldsymbol{\pi}$ and θ are shown in equation (4) where *i* denotes respondent, *r* denotes replication, *N* denotes the total number of respondents, *R* denotes the total number of replications, $\widehat{\pi_{i1r}}$ denotes the posterior sample mean of π_{i1} in replication *r* and $\widehat{\theta_{ir}}$ denotes the posterior sample mean of θ_i in replication *r*. A RMSE in this case measures the bias and variance of a posterior sample mean in a replication and the expected squared difference between a posterior sample mean in a replication and the true parameter is approximated by the average squared difference between a posterior sample mean in a replication and the true parameter across replications.

$$\text{Average RMSE}(\widehat{\pi_{i1r}}) = \frac{1}{N}\sum_{i=1}^{N}\text{RMSE}(\widehat{\pi_{i1r}}) \qquad (4)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sqrt{\text{MSE}(\widehat{\pi_{i1r}})}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sqrt{E[(\widehat{\pi_{i1r}} - \pi_{i1})^2]}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sqrt{\text{Bias}(\widehat{\pi_{i1r}})^2 + \text{Variance}\ (\widehat{\pi_{i1r}})}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\widehat{\pi_{i1r}} - \pi_{i1})^2}$$

$$\text{Average RMSE}(\widehat{\theta_{ir}}) = \frac{1}{N}\sum_{i=1}^{N}\text{RMSE}(\widehat{\theta_{ir}})$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sqrt{\text{MSE}(\widehat{\theta_{ir}})}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sqrt{E\left[(\widehat{\theta_{ir}} - \theta_i)^2\right]}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sqrt{\text{Bias}(\widehat{\theta_{ir}})^2 + \text{Variance}\,(\widehat{\theta_{ir}})}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\widehat{\theta_{ir}} - \theta_i)^2}$$

The bias of an estimator is the difference between the expected value of the estimator and the true parameter. In this study, the bias of $\widehat{\pi_{i1r}}$ and $\widehat{\theta_{ir}}$ are averaged over respondents respectively and the corresponding mathematical definitions are shown in equation (5). The expected values of $\widehat{\pi_{i1r}}$ and $\widehat{\theta_{ir}}$ are approximated by the average $\widehat{\pi_{i1r}}$ and the average $\widehat{\theta_{ir}}$ across replications respectively.

$$\text{Average Bias}(\widehat{\pi_{i1r}}) = \frac{1}{N}\sum_{i=1}^{N}(E[\widehat{\pi_{i1r}}] - \pi_{i1}) \qquad (5)$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{R}\sum_{r=1}^{R}\widehat{\pi_{i1r}} - \pi_{i1}\right)$$

$$\text{Average Bias}(\widehat{\theta_{ir}}) = \frac{1}{N}\sum_{i=1}^{N}\left(E\left[\widehat{\theta_{ir}}\right] - \theta_i\right)$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{R}\sum_{r=1}^{R}\widehat{\theta_{ir}} - \theta_i\right)$$

An example of the relationship between $\widehat{\pi_{i1r}}$ and π_{i1} over respondents when N = 300 and J = 15 is shown in Figure 2. Since this relationship is not best described as a linear relationship and such pattern is observed across conditions, to report the correlation between $\widehat{\pi_{i1r}}$ and π_{i1}, Kendall's τ coefficient is calculated for each iteration and the coefficients are averaged across replications to obtain an average $\tau(\widehat{\pi_{i1r}}, \pi_{i1})$ as shown in equation (6). Figure 2 also shows an example of the relationship between $\widehat{\theta_{ir}}$ and θ_i over respondents when N = 300 and J = 15. Since

this relationship appears to be linear and such pattern is observed across conditions, Pearson correlation coefficient is calculated to examine the correlation between $\widehat{\theta_{ir}}$ and θ_i.

$$\text{Average } \tau(\widehat{\pi_{i1r}}, \pi_{i1}) = \frac{1}{R}\sum_{r=1}^{R} \tau(\widehat{\pi_{i1r}}, \pi_{i1}) \quad (6)$$

$$\text{Average } \rho(\widehat{\theta_{ir}}, \theta_i) = \frac{1}{R}\sum_{r=1}^{R} \rho(\widehat{\theta_{ir}}, \theta_i)$$

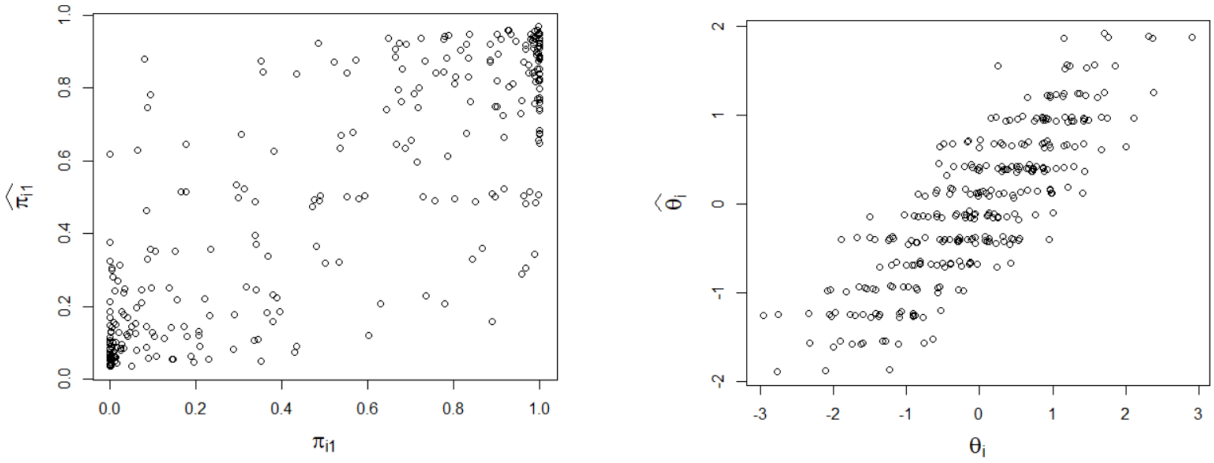

Figure 2. The left panel shows the relationship between $\widehat{\pi_{i1r}}$ and π_{i1} over respondents. The right panel shows the relationship between $\widehat{\theta_{ir}}$ and θ_i over respondents. The estimates of the parameters in both of the panels were obtained in one of the replications when N = 300 and J = 15.

The recovery of z_{ij} is evaluated using an average PCR. A PCR of a respondent is defined as the proportion of items for which the estimate of z_{ij} is the same as the true z_{ij}. As shown in equation (7), an average PCR is calculated by averaging the PCR's over respondents and over replications. Given that z_{ij} is a categorical variable, $\widehat{z_{ijr}}$ in equation (6) is the posterior sample

mode rather than the posterior sample mean of z_{ij} in replication *r*. The larger the average PCR($\widehat{z_{ijr}}$), the better the z_{ij}'s are considered to be recovered.

$$\text{Average PCR}(\widehat{z_{ijr}}) = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\sum_{j=1}^{J} I(\widehat{z_{ijr}} = z_{ij})}{J}\right)\right) \qquad (7)$$

4.4 Results

In each MCMC chain, the starting values for θ_i were generated from *Normal(0,* 3^2*)* and the starting values for Z_{ij} were generated from *Multinomial(0.5, 0.5).* In this study, since the item difficulty parameters were known, label switching was unlikely to happen. However, in order to be certain, label switching was still checked. No label switching within an MCMC chain or across replications was observed in the posterior samples. Convergence of the MCMC chains was diagnosed using the Potential Scale Reduction Factor (PSRF; Gelman et al., 2013). After a burn-in of 10,000 iterations, the PSRF remained very close to 1 and smaller than 1.1, which is usually considered as a sign that an MCMC chain has converged (Sinharay, 2003). After the burn-in, each MCMC chain continued to run another 10,000 iterations.

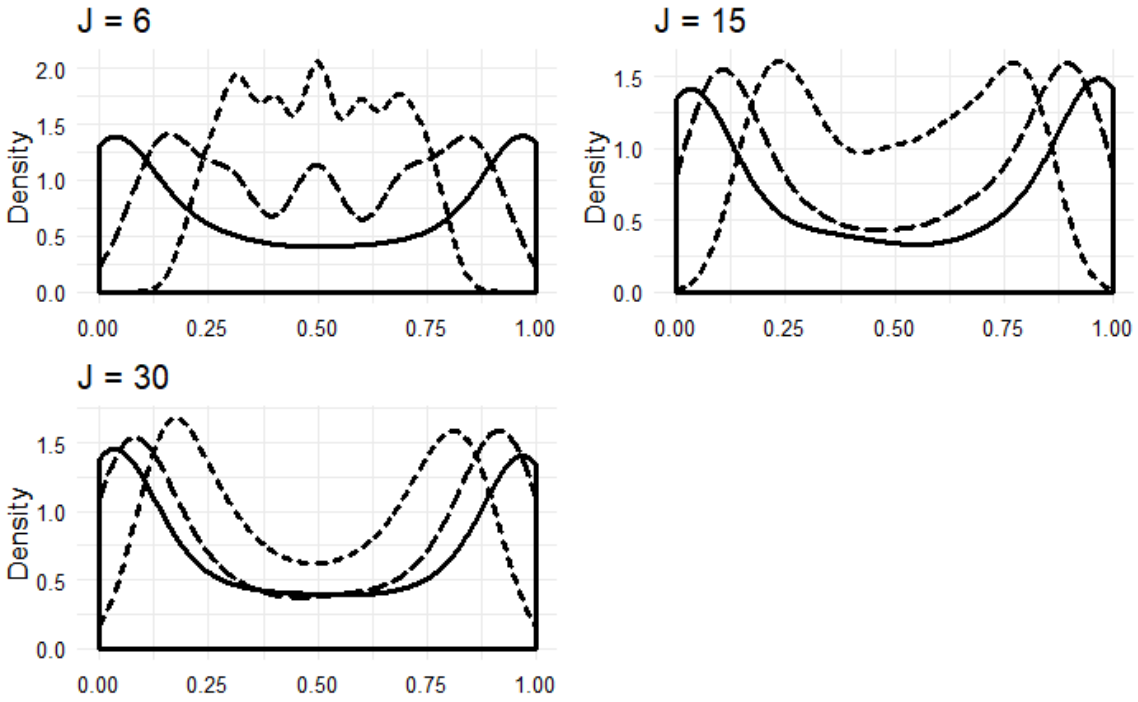

Figure 3. The distribution of $\widehat{\pi_{i1}}$ and π_{i1} across respondents in one of the replications for N = 1000 in Study I. The x-axis in each panel is either $\widehat{\pi_{i1}}$ or π_{i1}. The solid curve shows the distribution of the π_{i1} across respondents. The dashed curve shows the distribution of $\widehat{\pi_{i1}}$ across respondents when the prior on $\boldsymbol{\pi_i}$ was *Dirichlet*(0.25, 0.25). The dotted curve shows the distribution of $\widehat{\pi_{i1}}$ across respondents when the prior for $\boldsymbol{\pi_i}$ was *Dirichlet*(1, 1). Different panels are for a different test length J.

Table 3. Parameter recovery in Study I evaluated by average PCR, average RMSE, average bias and average correlation

N	P($\boldsymbol{\pi_i}$)	J	$\widehat{z_{ijr}}$	$\widehat{\pi_{i1r}}$			$\widehat{\theta_{ir}}$		
			PCR	RMSE	Bias	τ	RMSE	Bias	ρ
300	*Dir*(0.25, 0.25)	6	0.77	0.29	0.00	0.45	0.68	0.07	0.70
		15	0.83	0.21	0.00	0.58	0.55	0.00	0.84
		30	0.87	0.15	0.00	0.66	0.41	0.00	0.91
	Dir(1, 1)	6	0.76	0.31	-0.01	0.46	0.67	0.07	0.70
		15	0.82	0.24	0.00	0.58	0.54	0.00	0.84
		30	0.86	0.18	-0.01	0.66	0.41	0.01	0.91
500	*Dir*(0.25, 0.25)	6	0.78	0.29	0.01	0.46	0.68	-0.03	0.67
		15	0.85	0.21	0.00	0.58	0.53	-0.01	0.83
		30	0.88	0.14	0.00	0.65	0.42	0.01	0.91
	Dir(1,1)	6	0.76	0.30	0.01	0.46	0.68	-0.02	0.67
		15	0.83	0.24	0.00	0.58	0.52	-0.01	0.83
		30	0.87	0.18	0.00	0.65	0.41	0.01	0.91
1000	*Dir*(0.25, 0.25)	6	0.77	0.29	0.00	0.46	0.69	0.01	0.70
		15	0.84	0.21	0.00	0.58	0.54	0.01	0.83
		30	0.87	0.15	0.01	0.66	0.42	-0.01	0.91
	Dir(1,1)	6	0.76	0.30	0.00	0.46	0.69	0.01	0.71
		15	0.83	0.24	-0.01	0.58	0.54	0.01	0.83
		30	0.86	0.18	0.01	0.66	0.42	0.00	0.91

Note. *Dir* is short for *Dirichlet*.

The parameter recovery statistics for Study I are reported in Table 3. Given a test length and a prior on $\boldsymbol{\pi_i}$, the average RMSE($\widehat{\pi_{i1r}}$) and the average PCR(z) do not seem to be influenced by sample size. For each prior on $\boldsymbol{\pi_i}$, the average RMSE($\widehat{\pi_{i1r}}$) decreases and the average PCR(z) increases as test length increases. For each test length, the average RMSE($\widehat{\pi_{i1r}}$) increases and the average PCR(z) decreases slightly for the flat prior on $\boldsymbol{\pi_i}$ compared with the true prior. However, the flat prior and a longer test still return a smaller average RMSE($\widehat{\pi_{i1r}}$) and a larger average PCR(z) compared with the true prior and a shorter test. For example, when N = 300, J = 30 and P($\boldsymbol{\pi_i}$) ~ *Dirichlet*(1, 1), the average RMSE($\widehat{\pi_{i1r}}$) is 0.18 and the average PCR(z) is 0.87, whereas when N = 300, J = 6 and P($\boldsymbol{\pi_i}$) ~ *Dirichlet*(0.25, 0.25), the average RMSE($\widehat{\pi_{i1r}}$) is 0.29 and the average PCR(z) is 0.77. The average Bias($\boldsymbol{\pi}$) under all the conditions is consistently small and

very close to 0. The average $\tau(\widehat{\pi_{i1r}}, \pi_{i1})$ increases as test length increases and does not seem to be influenced by sample size and prior choice for $\boldsymbol{\pi_i}$. Given a test length, the average RMSE($\widehat{\theta_{ir}}$) and the average $\rho(\widehat{\theta_{ir}}, \theta_i)$ do not seem to be influenced by sample size and prior choice for $\boldsymbol{\pi_i}$. The average RMSE($\widehat{\theta_{ir}}$) decreases and the average $\rho(\widehat{\theta_{ir}}, \theta_i)$ increases as test length increases. The average Bias($\widehat{\theta_{ir}}$**)** is the largest when both the sample size and the test length were the smallest. When sample size increases from 300 to 500 and when test length increases from 6 to 15, the average Bias($\widehat{\theta_{ir}}$**)** decreases and becomes closer to 0.

Prior choice for $\boldsymbol{\pi_i}$ and test length also affected the distribution of $\widehat{\pi_{i1r}}$ across respondents. When the test was longer and when the true prior was used for $\boldsymbol{\pi_i}$, the distribution of $\widehat{\pi_{i1r}}$ across respondents tended to better recover the distribution of π_{i1} across respondents compared with when the test was shorter and when the flat prior was used. In all the conditions, the distributions of $\widehat{\pi_{i1r}}$ tended to shrink towards the mean of the generating distribution of $\boldsymbol{\pi_i}$ which is 0.5. As an example, Figure 3 shows the distributions of π_{i1} and $\widehat{\pi_{i1r}}$ for one of the 50 replications when *N* = 1000.

CHAPTER 5

STUDY II

5.1 Purpose and Study Design

The purpose of study II was to investigate how well the parameters in the MMR could be recovered under varying conditions when item difficulty parameters for each latent group are unknown. This is the kind of situation that would occur, for example, for a newly created test. The design of the simulation study was the same as the one in Study I. The only difference between Study I and Study II was that item difficulty parameters were unknown and need to be estimated in Study II.

As was the case for Study I, the starting values of θ_i for the MCMC algorithm were generated from *Normal(0,* 3^2*)* and the starting values of Z_{ij} were generated from *Multinomial(0.5, 0.5).* The starting values of b_{gj} were generated from *Normal(0,* 1.5^2*)*. Each MCMC chain was run 20,000 iterations with the first 10,000 iterations used as a burn-in to ensure that the PSRF remained considerably smaller than 1.1 after the burn-in. As was observed in Study I, no label switching jumps were observed in the posterior samples within a single MCMC chain. Label switching that occurred across MCMC chains was corrected based on the correlation between the posterior estimates of item difficulty parameters and the generating parameters. Suppose in a certain replication, the posterior estimates in latent group 1 are positively correlated with the generating parameters in latent group 2. These posterior estimates are then relabeled as latent group 2. After the correction, latent group so that the labels became consistent across replications.

5.2 Evaluation Statistics for Parameter Recovery

As was done in Study I, the recoveries of membership probability $\boldsymbol{\pi_i}$ and ability θ_i were evaluated using the average RMSE, the average bias and the average correlation as shown in equations (4) – (6). The recovery of latent group membership Z_{ij} was evaluated using the average PCR as shown in equation (7). Because the item difficulty parameters were estimated in Study II, their recovery was also evaluated using the average RMSE and the average bias across items and latent groups and the average correlation across replications. The definition of the average RMSE for item difficulty is shown in equation (8) where *J* denotes the number of test items, *G* denotes the number of latent groups and $\widehat{b_{gJr}}$ denotes the posterior sample mean of b_{gj} in replication *r*. The definition of the average bias of $\widehat{b_{gJr}}$ is shown in equation (9) where the $E\left[\widehat{b_{gJr}}\right]$ is approximated by the average of $\widehat{b_{gJr}}$ across replications. The Pearson correlation coefficient between $\widehat{b_{gJr}}$ and b_{gj} over items and latent groups was calculated for each replication and the coefficients are averaged across replications to obtain an average correlation as shown in equation (10).

$$\text{Average RMSE}(\widehat{b_{gJr}}) = \frac{1}{G\times J}\sum_{g=1}^{G}\sum_{j=1}^{J}\text{RMSE}(\widehat{b_{gJr}}) \qquad (8)$$

$$= \frac{1}{G\times J}\sum_{g=1}^{G}\sum_{j=1}^{J}\sqrt{\text{MSE}(\widehat{b_{gJr}})}$$

$$= \frac{1}{G\times J}\sum_{g=1}^{G}\sum_{j=1}^{J}\sqrt{E\left[(\widehat{b_{gJr}} - b_{gj})^2\right]}$$

$$= \frac{1}{G\times J}\sum_{g=1}^{G}\sum_{j=1}^{J}\sqrt{\text{Bias}(\widehat{b_{gJr}})^2 + \text{Variance}\,(\widehat{b_{gJr}})}$$

$$\approx \frac{1}{G\times J}\sum_{g=1}^{G}\sum_{j=1}^{J}\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\widehat{b_{gJr}} - b_{gj})^2}$$

$$\text{Average Bias}(\widehat{b_{gjr}}) = \frac{1}{G\times J}\sum_{g=1}^{G}\sum_{j=1}^{J}\left(E\left[\widehat{b_{gjr}}\right] - b_{gj}\right) \qquad (9)$$

$$\approx \frac{1}{G\times J}\sum_{g=1}^{G}\sum_{j=1}^{J}\left(\frac{1}{R}\sum_{r=1}^{R}\widehat{b_{gjr}} - b_{gj}\right)$$

$$\text{Average } \rho(\widehat{b_s}, b) = \frac{1}{R}\sum_{r=1}^{R}\rho(\widehat{b_r}, b) \qquad (10)$$

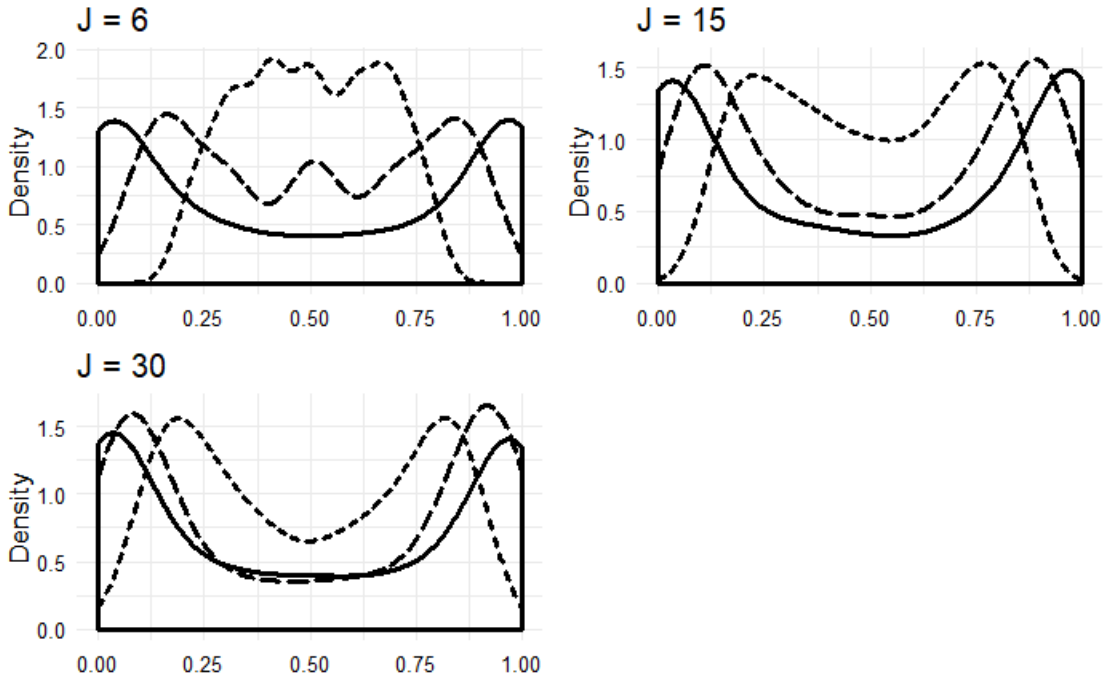

Figure 4. The distribution of $\widehat{\pi_{i1}}$ and π_{i1} across respondents in one of the replications for sample size *N* = 1000 in Study II. The x-axis in each panel is either $\widehat{\pi_{i1}}$ or π_{i1}. The solid curve shows the distribution of the π_{i1} across respondents. The dashed curve shows the distribution of $\widehat{\pi_{i1}}$ across respondents when *Dirichlet*(0.25, 0.25) was the prior on $\boldsymbol{\pi_i}$. The dotted curve shows the distribution of $\widehat{\pi_{i1}}$ across respondents when *Dirichlet*(1, 1) was the prior on $\boldsymbol{\pi_i}$. Different panels are for the three different test lengths *J*.

5.3. Results

The parameter recovery statistics for Study II are reported in Table 4. Given a prior on $\boldsymbol{\pi_i}$ and a test length, the average PCR(z) and the average RMSE($\widehat{\pi_{i1r}}$) do not seem to be influenced by sample size. For a given prior on $\boldsymbol{\pi_i}$, the average PCR(z) increases and the average RMSE($\widehat{\pi_{i1r}}$)

decreases as test length increases. For a given test length, the average PCR(z) decreases slightly and the average RMSE($\widehat{\pi_{i1r}}$) increases when the prior for $\boldsymbol{\pi_i}$ changes from *Dirichlet*(0.25, 0.25) to *Dirichlet*(1, 1). A larger test length, however, appears to reduce the negative effect of a flat prior on the average PCR(z) and the average RMSE($\widehat{\pi_{i1r}}$). For example, when N = 300, J = 30 and P($\boldsymbol{\pi_i}$) ~ *Dirichlet*(1, 1), the average PCR(z) is 0.85 and the average RMSE($\widehat{\pi_{i1r}}$) is 0.19, whereas when N = 300, J = 6 and P($\boldsymbol{\pi_i}$) ~ *Dirichlet*(0.25, 0.25), the average PCR(z) is 0.78 and the average RMSE($\widehat{\pi_{i1r}}$) is 0.29. The average Bias($\boldsymbol{\pi}$) is consistently negligible (i.e., close to zero) across all the conditions. The average $\tau(\widehat{\pi_{i1r}}, \pi_{i1})$ increases as test length increases and is not influenced by sample size and prior choice for $\boldsymbol{\pi_i}$.

For a given test length, the average RMSE($\widehat{b_{gJr}}$) decreases considerably as sample size increases when the true prior on $\boldsymbol{\pi_i}$ is used. For a given sample size and when the true prior on $\boldsymbol{\pi_i}$ is used, the average RMSE($\widehat{b_{gJr}}$) decreases slightly as test length increases from 6 to 15 items. Less of a decrease is evident as test length increases from 15 to 30 items. When the flat prior on $\boldsymbol{\pi_i}$ is used, the average RMSE($\widehat{b_{gJr}}$) tends to be larger than when the true prior is used for a given sample size and test length. Further, increasing test length and sample size do not seem to decrease the average RMSE($\widehat{b_{gJr}}$). The Average Bias($\widehat{b_{gJr}}$) is very close to 0 in all the conditions. The average $\rho(\hat{b}_s, b)$ is consistently positive and large in all the conditions.

The average RMSE($\widehat{\theta_{ir}}$) and the average $\rho(\widehat{\theta_{ir}}, \theta_i)$ seems to be only affected by test length. The average RMSE($\widehat{\theta_{ir}}$) decreases and the average $\rho(\widehat{\theta_{ir}}, \theta_i)$ increases as test length increases. When both sample size and test length are the smallest, the average Bias($\widehat{\theta_{ir}}$) tends to be relatively large. Increasing sample size or test length appears to help decrease the average Bias($\widehat{\theta_{ir}}$) to 0.

As was observed in Study I, test length and choice of choice on $\boldsymbol{\pi_i}$ also influences the distributions of $\widehat{\pi_{i1r}}$ across respondents in Study II. The distribution of $\widehat{\pi_{i1r}}$ across respondents better approximated the distribution of π_{i1} across respondents when the true prior rather than the flat prior for $\boldsymbol{\pi_i}$ was used and for the longer test lengths. An example of such patterns is shown in Figure 4.

Table 4. Parameter recovery in study II evaluated by average PCR, average RMSE, average bias, and average correlation

N	P($\boldsymbol{\pi_i}$)	*J*	$\widehat{z_{ijr}}$	$\widehat{\pi_{i1r}}$			$\widehat{b_{gjr}}$			$\widehat{\theta_{ir}}$		
			PCR	RMSE	Bias	τ	RMSE	Bias	ρ	RMSE	Bias	ρ
300	*Dir*(0.25, 0.25)	6	0.78	0.29	-0.01	0.46	0.28	0.00	0.98	0.68	0.07	0.70
		15	0.83	0.21	-0.00	0.57	0.25	0.00	0.98	0.55	0.00	0.84
		30	0.87	0.15	-0.01	0.65	0.24	0.00	0.98	0.41	0.00	0.91
	Dir(1, 1)	6	0.76	0.31	-0.01	0.46	0.26	0.00	0.98	0.67	0.07	0.70
		15	0.81	0.24	-0.00	0.58	0.28	0.00	0.98	0.55	0.00	0.84
		30	0.85	0.19	-0.02	0.65	0.28	0.00	0.98	0.42	0.00	0.91
500	*Dir*(0.25, 0.25)	6	0.78	0.29	0.01	0.46	0.20	0.00	0.99	0.68	-0.02	0.68
		15	0.84	0.21	0.00	0.58	0.18	0.00	0.99	0.53	0.00	0.83
		30	0.88	0.14	0.00	0.65	0.18	0.00	0.99	0.41	0.01	0.91
	Dir(1,1)	6	0.75	0.30	0.01	0.45	0.28	0.00	0.99	0.68	-0.02	0.68
		15	0.82	0.24	-0.00	0.58	0.32	0.00	0.99	0.53	0.00	0.83
		30	0.85	0.19	0.01	0.65	0.32	0.00	0.99	0.42	0.01	0.91
1000	*Dir*(0.25, 0.25)	6	0.78	0.29	-0.00	0.46	0.15	0.00	0.99	0.69	0.01	0.70
		15	0.84	0.21	-0.01	0.58	0.14	0.00	0.99	0.54	0.01	0.83
		30	0.87	0.15	0.00	0.66	0.13	0.00	0.99	0.42	-0.01	0.91
	Dir(1,1)	6	0.75	0.30	0.00	0.45	0.21	0.00	0.99	0.69	0.01	0.71
		15	0.81	0.24	-0.01	0.58	0.32	0.00	0.99	0.54	0.00	0.83
		30	0.85	0.19	0.01	0.66	0.33	0.00	0.99	0.42	-0.01	0.91

CHAPTER 6

STUDY III

6.1 Purpose and Study Design

In order to fit an MMR, the number of latent groups has to be specified. Selection of the best fitting model, i.e., the model with the appropriate number of latent groups, is critical to ensure that the model does not under-fit or over-fit the data. Likelihood-based goodness-of-fit measures are one of the most common approaches for model selection and usually evaluate the tradeoff between log-likelihood and model complexity. Previous studies on mixed membership models and mixture IRT models (Erosheva, 2002; Erosheva & Fineberg, 2007; Gormley & Murphy, 2009; Li, et al., 2009) employed Akaike's information criterion (AIC; Akaike, 1974), Bayesian Information Criterion (BIC; Schwartz, 1978), deviance information criterion (DIC; Spiegelhalter, Best, Garlin & Van Der Linde, 2002), and Akaike's information criterion for MCMC samples (AICM; Raftery et al., 2007) to determine the number of latent groups.

AIC and BIC were originally developed for maximum likelihood estimators and were modified for use with MCMC estimation (Congdon, 2003). According to Congdon (2003), the modified AIC and BIC for MCMC estimation are defined as follows:

$$\text{AIC} = \overline{D(\boldsymbol{\delta})} + 2p \qquad (9)$$

$$\text{BIC} = \overline{D(\boldsymbol{\delta})} + log(N)p$$

$$\overline{D(\boldsymbol{\delta})} = -2\frac{1}{S}\sum_{s=1}^{S} l(\boldsymbol{\delta}^{(s)}|X)$$

where $\overline{D(\boldsymbol{\delta})}$ is the posterior mean deviance given relevant parameters $\boldsymbol{\delta}$, p is the number of estimated parameters, N is the number of observations, and $l(\boldsymbol{\delta}^{(s)}|X)$ is the log-likelihood in

iteration *s* after burn-in. The definition of *p* and *N*, however, is not straightforward in Bayesian models especially in those that have informative prior information (Erosheva, 2002; Spiegelhalter, Best, Carlin & van der Linde, 2014).

DIC can be conveniently computed with MCMC estimation and is defined as follows:

$$\text{DIC} = \text{D}(\bar{\boldsymbol{\delta}}) + 2{*}p_D \qquad (10)$$

$$p_D = \overline{D(\boldsymbol{\delta})} - \text{D}(\bar{\boldsymbol{\delta}})$$

$$\text{D}(\bar{\boldsymbol{\delta}}) = -2l(\hat{\boldsymbol{\delta}}|X)$$

where $\text{D}(\bar{\boldsymbol{\delta}})$ is the deviance given the point estimates of the parameters calculated using posterior samples. DIC has been criticized for its tendency to overfit, that is, to select more complex models (Van der Linde, 2005, 2012). Thus, Plummer (2008) and Ando (2012) recommended DIC^* that has a larger penalty on model complexity than DIC.

$$\text{DIC}^* = \text{D}(\bar{\boldsymbol{\delta}}) + 3{*}p_D \qquad (11)$$

For AIC, BIC, DIC and DIC^*, the models with smaller values among candidate models are preferred.

AICM is analogous to AIC and was developed for use with MCMC estimation. It is defined as:

$$\text{AICM} = 2(\overline{l(\boldsymbol{\delta}|X)} - s^2_{l(\boldsymbol{\delta}|X)}) \qquad (12)$$

$$\overline{l(\boldsymbol{\delta}|X)} = \frac{1}{S}\sum_{s=1}^{S} l(\boldsymbol{\delta}^{(s)}|X)$$

$$s^2_{l(\boldsymbol{\delta}|X} = \frac{1}{S}\sum_{s=1}^{S}(l\left(\boldsymbol{\delta}^{(s)}\middle|X\right) - \overline{l(\boldsymbol{\delta}|X)})^2$$

The models with larger AICM are preferred.

Previous studies showed that AIC, BIC, DIC and AICM may not always agree on the best fitting models with latent groups and some of them outperform the others in certain conditions for certain models (Erosheva & Fineberg, 2007; Li, et al., 2009). For example, the

BIC, modified for use with MCMC estimation, has been found to outperform the other measures for mixture IRT models (Li, et al., 2009), whereas AICM has been found to be one of the best-performing model selection measures in a mixed membership model (Erosheva & Fineberg, 2007).

In this chapter, the performance of AIC, BIC, DIC, DIC* and AICM for selecting the best fitting MMR model estimated using MCMC under different testing conditions was examined using simulated data. The calculations of these measures follow the above equations. In equation (9), p is defined as the total number of item difficulty parameters across latent groups, which is consistent with the definition in Li, e. al.'s study (2009). For example, in a two-latent-group MMR model that is fitted to responses to 15 test items, p is equal to $15 \times 2 = 30$. In equation (10), $l(\widehat{\boldsymbol{\delta}}|X)$ is defined as

$$l(\widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{b}}|\boldsymbol{X}) = \log\Big(\prod_{i=1}^{N} \prod_{j=1}^{J} \sum_{g=1}^{G} \widehat{\pi_{ig}}^{s} P_g\big(X_{ij}\big|\widehat{Z_{ijg}},\ \widehat{\theta_i},\ \widehat{b_{gj}}\big)\Big)$$

where $\widehat{\pi_{ig}}$, $\widehat{\theta_i}$, and $\widehat{b_{gj}}$ are posterior sample means and $\widehat{Z_{ijg}}$ is a posterior sample mode. In equation (12), $l(\boldsymbol{\delta}^{(s)}|X)$ is defined as

$$l(\boldsymbol{\pi}^s, \boldsymbol{\theta}^s, \boldsymbol{b}^s|\boldsymbol{X}) = \log\Big(\prod_{i=1}^{N} \prod_{j=1}^{J} \sum_{g=1}^{G} \pi_{ig}{}^{s} P_g\big(X_{ij}\big|Z_{ijg}{}^{s},\ \theta_i{}^{s}, b_{gj}{}^{s}\big)\Big)$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{J} log\Big\{\sum_{g=1}^{G} \pi_{ig}{}^{s} P_g\big(X_{ij}\big|Z_{ijg}{}^{s},\ \theta_i{}^{s}, b_{gj}{}^{s}\big)\Big\}.$$

The simulation conditions for this study were the same as those described for Study I. That is, the data were generated given two latent groups, three sample sizes (N = 300, 500 and 1000) and three test lengths (J = 6, 15 and 30 items). For each condition, the simulation was replicated 50 times. For each simulated data set, the MMR models with no latent groups, two latent groups, three latent groups and four latent groups were fitted using the MCMC algorithm. The *Dirichlet*(0.25, 0.25) prior on $\boldsymbol{\pi_i}$ is the same as the generating distribution of $\boldsymbol{\pi_i}$. Each MCMC chain was run for 20,000 iterations with the first 10,000 discarded as a burn-in.

6.2. Results

Table 5 shows how often an information criterion measure picks a model as the optimal model across the 50 replications under each testing condition. BIC appears to be the best-performing measure since it picks the models with two latent groups 100% of the time across all conditions. AIC is next closest in selecting the correct model although its performance is inconsistent across testing conditions. In some of the conditions, AIC selects the correct model more than 94% of the time whereas in other conditions, AIC fails to capture the correct model at all. Moreover, AIC tends to pick more complex models. DIC and DIC* consistently miss the correct model across all the conditions and favor more complex models. AICM is close to missing the correct model across all the conditions all the time. In some conditions, it favores the simplest model and in some other conditions, it favors the most complex model.

Table 5. Percent of replications in which an information criterion picked out a model as the optimal model.

N	*J*	*G*	AIC	BIC	DIC	DIC*	AICM
300	6	1	0	0	0	0	96
		2	**98**	**100**	**0**	**0**	**4**
		3	2	0	28	20	0
		4	0	0	72	80	0
	15	1	0	0	0	0	6
		2	**70**	**100**	**0**	**0**	**0**
		3	26	0	18	20	6
		4	4	0	82	80	88
	30	1	0	0	0	10	100
		2	**20**	**100**	**0**	**0**	**0**
		3	70	0	6	0	0
		4	10	0	94	90	0
500	6	1	0	0	0	0	0
		2	**84**	**100**	**0**	**0**	**0**
		3	16	0	20	20	28
		4	0	0	80	80	72
	15	1	0	0	0	0	44
		2	**20**	**100**	**0**	**0**	**0**
		3	48	0	2	2	0
		4	32	0	98	98	56
	30	1	0	0	0	0	100
		2	**0**	**100**	**0**	**0**	**0**
		3	4	0	2	4	0
		4	96	0	98	96	0
1000	6	1	0	0	0	0	0
		2	**94**	**100**	**0**	**0**	**0**
		3	6	0	40	28	36
		4	0	0	60	72	64
	15	1	0	0	0	0	64
		2	**0**	**100**	**0**	**0**	**0**
		3	0	0	0	0	0
		4	100	0	100	100	36
	30	1	0	0	0	0	100
		2	**0**	**100**	**0**	**0**	**0**
		3	0	0	0	0	0
		4	100	0	100	100	0

Note. The models with 2 latent groups are the true models.

CHAPTER 7

SUMMARY AND DISCUSSION

IRT models have been widely applied to analyze educational test data. They are usually used to estimate respondents' ability as well as the characteristics of test items such as item difficulty and item discrimination. The Rasch model is the simplest IRT model. It only includes item difficulty for the item characteristics. Item discrimination is represented in the Rasch model but is typically assumed to be 1 for all items.

One of the variants of the conventional IRT models is mixture IRT. A mixture IRT model accounts for the possible heterogeneity in model parameters by detecting latent groups of respondents in the data. For example, respondents may solve problems using different strategies and may show different test-taking behaviors. The same set of test items may perform differently for different groups of respondents.

Mixture IRT models can be used to accommodate this scenario and to estimate respondents' latent group membership as well as the item characteristics under each latent group. The assumption in mixture IRT models is that each respondent remains in the same latent group across the entire test. Using the analogy above of different problem-solving strategies, a mixture IRT model assumes that the respondents in a given latent group use the same problem-solving strategy over all test items. It is possible, however, that a respondent switches between problem-solving strategies across items. To account for this possibility, in this dissertation, a mixed membership Rasch (MMR) model was developed by integrating the Rasch model into the framework of mixed membership models.

The MMR allows a respondent to belong to all the possible latent groups in the data but with different probabilities at the test level. This is represented by a membership probability vector $\boldsymbol{\pi_i}$ for each respondent. At the item level, the MMR assumes that a respondent belongs to one of the latent groups for a given item, denoted by Z_{ij}, and the respondent may belong to different latent groups for different items. For a given latent group, the probability of a correct response to an item is parameterized using the Rasch model.

In the Rasch model as well as for IRT models in general, the scale of item difficulty and ability parameters is undetermined. This can potentially result in nonconvergence of an estimation algorithm for the IRT models. Therefore, researchers and IRT analysis software usually implement a scale on either item difficulty or ability parameters. For mixture IRT and the MMR, it is important to not only determine a scale for the parameters but also to make sure that the parameters across latent groups are on the same scale given how the parameter estimates are sampled iteratively in the estimation algorithms as well as the need that the parameters across latent groups are comparable for interpretability purposes. Consistent with common practice in studies employing a mixture Rasch model, a constraint that the item difficulty parameters sum to 0 as described in Rost (1990) was imposed for each latent group. In this way, the parameters in the Rasch model had a fixed scale and were on the same scale across latent groups. This constraint was considered reasonable in the current simulation studies since the sum of the generating item difficulty parameters was 0 for each latent group. However, when the sum of the generating item difficulty parameters is not 0, such a constraint would bias the estimates and more appropriate scaling methods would need to be employed (Paek & Cho, 2015).

In this dissertation, simulation studies were conducted to investigate how well the parameters in the MMR could be recovered under practical testing conditions and what factors

affected the recovery. The MMR was estimated using a Metropolis-within-Gibbs algorithm. The results of the simulation studies showed that the recovery of $\boldsymbol{\pi_i}$ and Z_{ij} improved as test length increased. The use of an incorrect prior for $\boldsymbol{\pi_i}$, that is, a prior that was different from the generating distribution of $\boldsymbol{\pi_i}$, negatively affected their recovery. Increasing test length appeared to be able to overcome such negative effects. Sample size, however, did not appear to influence the recovery of neither $\boldsymbol{\pi_i}$ nor Z_{ij}. The recovery of Z_{ij} was reasonably good even though there was only one observed data point that could be used to estimate Z_{ij}. The recovery of item difficulty parameters b_{gj} improved as sample size increased. The recovery also improved slightly as test length increased from small to medium. When the wrong prior was used for $\boldsymbol{\pi_i}$, the recovery of b_{gj} was worse compared with when the true prior was used. Increasing test length and sample size, however, did not seem to cancel out such negative effects. The recovery of ability parameters improved as test length increased but did not seem to be influenced by neither sample size nor choice of prior for $\boldsymbol{\pi_i}$.

An unexpected result was that knowing item difficulty parameters did not appear to improve the recovery of $\boldsymbol{\pi_i}$ and Z_{ij}. As shown in Chapter 3, the posterior of $\boldsymbol{\pi_i}$ is a function of Z_{ij} and the likelihood in the posterior of Z_{ij} is a function of ability and item difficulty parameters. Galyardt (2012) suggested that knowing the parameters in the likelihood of the posterior of Z_{ij} returned better recovery of $\boldsymbol{\pi_i}$ in a mixed membership model. In the current MMR, since the likelihood in the posterior of Z_{ij} is a function of both item difficulty and ability parameters, knowing only item difficulty parameters may not have been sufficient to improve the estimation of Z_{ij} and $\boldsymbol{\pi_i}$.

In this dissertation, parameter recovery was evaluated using RMSE, bias and the correlation between posterior estimates and generating parameters. For item difficulty and ability

parameters, it would be useful to compare their recovery obtained in the current simulation studies with previous simulation studies on IRT models to get some further insight as to how well the parameters in IRT models can be estimated in a model as complicated as the MMR. Such comparisons, however, are somewhat complicated as different studies have used different simulation designs and have defined recovery statistics in different ways even though the statistics are called by the same name (e.g., Natesan, Minka & Rubright, 2016; Si & Schumacker, 2004). Future research addressing these differences in recovery studies would be helpful.

In this dissertation, simulation studies were also conducted to evaluate how well different likelihood-based goodness-of-fit measures performed in selecting the optimal number of latent groups for the MMR. Results suggested that BIC consistently selected the model with the correct number of latent groups across different sample sizes and test lengths that were examined in this study. AIC's performance was less consistent across testing conditions and did not appear to improve as either sample size or test length increased. DIC and DIC* tended to consistently favor more complex models across testing conditions. AICM tended to favor the simplest model in some testing conditions and the most complex model in the other conditions. The patterns of AIC, BIC, DIC and DIC* were somewhat consistent with previous research by Li et al. (2009) on the performance of model selection indices for mixture IRT models. AICM has not been applied to select the number of latent groups for mixture IRT models. For mixed membership models, AICM has showed acceptable performance (Erosheva et al., 2007; Kim, 2019) For the MMR, however, AICM generally failed to select the correct model consistently across conditions. Such inconsistency in the performance of AICM across mixed membership models and their variants may indicate that the behavior of AICM is sensitive to how a model is set up.

It is also noted that even though BIC performed well in the current simulation study, both AIC and BIC are functions of the number of estimated parameters and determining the number of estimated parameters in hierarchical models such as the MMR and mixture IRT models estimated using Bayesian methods is not necessarily straightforward. Previous studies that involved using information criterion measures to select number of latent groups for Bayesian mixture IRT (Li et al., 2009) and mixed membership models (Erosheva et al, 2007) did not fully investigate this issue and many of the previous studies did not report how the number of estimated parameters was defined (e.g., Erosheva et al., 2007; Huang, 2016).

Overall, the results of this dissertation suggest that the MMR can be estimated well under certain conditions and thus has the potential to help researchers understand respondents' partial memberships in all the latent groups that might exist in the data and their behaviors of switching between latent groups across test items.

In this dissertation, how well the parameters in the MMR can be recovered under different conditions was examined only when there were two latent groups in the data. Future studies should increase the number of latent groups so that a more complete picture of the patterns of parameter recovery in different conditions can be drawn. Galyardt (2012) and Erosheva et al. (2007) showed that each mixed membership model could be rewritten as a finite mixture model with far more latent groups than the mixed membership model. Since mixture Rasch model follows the framework of finite mixture models, future studies may also focus on understanding the relationship between the MMR and the mixture Rasch models. This might help researchers further investigate the statistical properties of the MMR and the possible bias in the estimates of item difficulty and ability parameters in the mixture Rasch models when partial membership and switching behaviors should not be omitted.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of machine learning research*, *9*(Sep), 1981-2014.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139-160.

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716-723

Ando, T. (2012) Predictive Bayesian model selection. *Am. J. Math. Management Sci.*, *31*, 13–38.

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

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17-35.

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

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

Choi, I. H., & Wilson, M. (2015). Multidimensional classification of examinees using the mixture random weights linear logistic test model. *Educational and psychological measurement*, *75*(1), 78-101.

Congdon, P. (2003). *Applied Bayesian modelling*. New York: John Wiley.

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

Embretson, S. E. (2004). Application of two IRT models for construct validation to issues about spatial ability. *Metodologia de las Ciencias del Comportamiento, 5*(2), 159-180.

Embretson, S. E. (2007). Mixed Rasch models for measurement in cognitive psychology. In *Multivariate and mixture distribution Rasch models* (pp. 235-253). Springer, New York, NY.

Erosheva, E. A. (2002). Grade of membership and latent structure models with application to disability survey data. *Unpublished doctoral dissertation, Department of Statistics, Carnegie Mellon University*.

Erosheva, E. A., Fienberg, S. E., & Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The annals of applied statistics, 1*(2), 346.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5220-5227.

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

French, J. W. (1965). The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement, 25*(1), 9-28.

Galyardt, A. (2012). *Mixed membership distributions with applications to modeling multiple strategy usage* (Doctoral dissertation, Carnegie Mellon University).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gelman, A., & Rubin, D.B. (1992b). Inference from iterative simulation using multiple *sequences, Statistical Science, 45*, 457-511.

Gormley, I. C., & Murphy, T. B. (2009). A grade of membership model for rank data. *Bayesian Analysis, 4*(2), 265-295.

Huang, H. Y. (2016). Mixture random-effect IRT models for controlling extreme response style on rating scales. *Frontiers in psychology*, *7*, 1706.

Johns, J., & Woolf, B. (2006, July). A dynamic mixture model to detect student motivation and proficiency. In *Proceedings of the national conference on artificial intelligence* (Vol. 21, No. 1, p. 163). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Kim, S. (2019). *Constructed Response Data Analysis Using Structural Equation Modeling and Topic Modeling* (Doctoral dissertation, University of Georgia).

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer Science & Business Media.

Kyllonen, P. C., Lohman, D. F., & Snow, R. E. (1984). Effects of aptitudes, strategy training, and task facets on spatial task performance. *Journal of Educational Psychology, 76*(1), 130.

Lohman, D. F. (1979). *Spatial Ability: A Review and Reanalysis of the Correlational Literature* (Tech. Rep. No. 8). Stanford, CA: Stanford university, School of Education.

Li, F., Cohen, A. S., Kim, S. H., & Cho, S. J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*(5), 353-373.

Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: application of the modified Multilevel Mixture IRT model. *Frontiers in psychology*, *9*, 1372.

Lord, F. M., & Novick, M. R. (2008*). Statistical theories of mental test scores*. IAP.

Maia, I., Severo, M., & Santos, A. C. (2020). Application of the mixture item response theory model to the Self-Administered Food Security Survey Module for Children. *PloS one*, *15*(1), e0228099.

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

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

Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. D. (2016). Bayesian prior choice in IRT estimation using MCMC and variational Bayes. *Frontiers in psychology*, *7*, 1422.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31,* 200-219.

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

Plummer, M. (2008) Penalized loss functions for Bayesian model comparison. *Biostatistics, 9,* 523–539.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959

A.E. Raftery, M.A. Newton, J.M. Satagopan, P.N. Krivitsly. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics*, 8:1-45, Oxford University Press, Oxford, 2007.

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 (statistical methodology), 59*(4), 731-792.

Rittle-Johnson, B., & Siegler, R. S. (1999). Learning to spell: Variability, choice, and change in children's strategy use. *Child development*, *70*(2), 332-348.

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

Salomon, G. (1974). Internalization of filmic schematic operations in interaction with learners' aptitudes. *Journal of Educational Psychology, 66*(4), 499–511.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Schlosser, A., Neeman, Z., & Attali, Y. (2019). Differential performance in high versus low stakes tests: evidence from the GRE test. *The Economic Journal*, *129*(623), 2916-2948.

Sen, S., Cohen, A. S., & Kim, S. H. (2016). The impact of non-normality on extraction of spurious latent classes in mixture IRT models. *Applied psychological measurement*, *40*(2), 98-113.

Si, C. F., & Schumacker, R. E. (2004). Ability estimation under different item parameterization and scoring models. *International Journal of Testing*, *4*(2), 137-181.

Sinharay, S. (2003). Assessing convergence of the Markov chain Monte Carlo algorithms: A review. *ETS Research Report Series*, 2003(1), i-52.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116*(3), 250.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology), 64*(4), 583-639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 485-493.

Swanson, M. R. (2015). *Extending an IRT mixture model to detect random responders on non-cognitive polytomously scored assessments* (Doctoral dissertation, James Madison University).

Van Der Linde, A. (2005). DIC in variable selection. *Statistica Neerlandica, 59*(1), 45-56.

van der Linde, A. (2012). A Bayesian view of model complexity. *Statistica Neerlandica, 66*(3), 253-271.

Wang, Y. S., & Erosheva, E. A. (2015). Fitting mixed membership models using mixedmem. https://cran.r-project.org/web/packages/mixedMem/vignettes/mixedMem.pdf

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education, 22*(2), 185-205.

Woodbury, M. A., Clive, J., & Garson Jr, A. (1978). Mathematical typology: a grade of membership technique for obtaining disease definition. *Computers and biomedical research, 11*(3), 277-298.

Wu, M., & Adams, R. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics education research journal, 18*(2), 93-113.

APPENDIX A

R code for Study I

```
# R code for a Metropolis-within-Gibbs algorithm for the MMR (item difficulties are known)

# note: I tried to make sure that the object names in the following R code are somewhat
#consistent with the notations in the MMR to make reading the R code easier.

# alpha = the hyperparameter in the prior distribution of pi
# start.z = starting values of z
# start.theta = starting values of theta
# n.group = number of latent groups
# n.sample = number of individuals
# n.item = number of items
# n.iter = number of iterations
# data = binary response data matrix with rows as individuals and columns as items


library(gtools)


# create functions used to update pi.i (Gibbs sampler)
posterior.pi.i= function(z.i, n.group){
  alpha.new= alpha + table(factor(z.i, levels= 1:n.group))
  draw= rdirichlet(1, alpha.new)  # draw a sample from the approximate posterior distribution
of pi.i
  return(draw)
}


# create functions used to update z.ij (Gibbs sampler)
posterior.z.ij= function( pi.i, theta.i, b.j, x.ij){
  p.j= exp(theta.i- b.j)/ (1+ exp(theta.i- b.j))
  if (x.ij== 1){
    p.new= p.j*pi.i
    draw= rmultinom(1, 1, p.new)  # draw a sample from the approximate posterior distribution
of z.ij  when x.ij = 1
    z.tmp= which(draw==1)
  }
  else if (x.ij==0){
    p.new.2= (1-p.j)*pi.i
    draw= rmultinom(1,1, p.new.2)  # draw a sample from the approximate posterior
distribution of z.ij when x.ij = 0
    z.tmp= which(draw==1)
  }
```

```
  return(z.tmp)
}


# create functions used to update theta.i (Metropolis step)
likelihood.theta.i= function(z.i, theta.i, b, n.item, x.i){ # calculate the log-likelihood of
individual i's response to each item
  l= rep(NA, n.item)
  for (i in 1:n.item){
    b.gj= b[,i][z.i[i]]
    p.jg= exp(theta.i- b.gj)/ (1+ exp(theta.i- b.gj))
    l[i]= dbinom(x.i[i], size=1, prob= p.jg, log= T)
  }
  return(sum(l))
}
prior.theta.i= function(theta.i){ # evaluate the prior density function of theta at a given sample
of theta
  return(dnorm(theta.i, 0, 1, log= T))
}
posterior.theta.i= function(z.i, theta.i, b, n.item, x.i){ # calculate the posterior distribution on
log scale
  post= likelihood.theta.i (z.i, theta.i, b, n.item, x.i)+ prior.theta.i(theta.i)
  return(post)
}
proposal.theta.i= function(theta.i){ # draw a sample of theta from a proposal distribution
  draw= rnorm(1, theta.i, sd= 0.5)
  return(draw)
}


# MCMC.
run.mcmc = function(alpha, start.z, b, start.theta, n.group, n.sample, n.item, n.iter, data){

   # create objects to store MCMC samples
   pi= vector('list', n.iter-1)   # create a list to store samples of pi obtained in all the iterations.
   z= vector('list', n.iter)    # create a list to store samples of z obtained in all the iterations.
   theta= matrix(NA, nrow= n.iter, ncol= n.sample)    # create a matrix to stores samples of
theta obtained in all the iterations with rows as iterations and columns as individuals.
   z[[1]]= start.z      # store starting values of z
   theta[1,]= start.theta   # store starting values of theta

  # run MCMC
  for ( i in 2:n.iter){  # n.iter = number of iterations
    ### Update pi
    pi.mat= matrix(NA, nrow= n.sample, ncol=n.group)     # In each iteration, samples are
stored in a matrix with rows as individuals and columns as groups.
```

```
    for (j in 1:n.sample){
      pi.mat[j,]= posterior.pi.i(z.i= z[[i-1]][j,], n.group)
    }
    pi[[i-1]]= pi.mat

    ### update z
    z.mat= matrix(NA, nrow= n.sample, ncol= n.item)  # In each iteration, samples are stored in
a matrix with rows as individuals and columns as items.
    for (j in 1:n.sample){
      for (n in 1:n.item){
        z.mat[j, n]= posterior.z.ij(pi[[i-1]][j,], theta[i-1, j], b[, n], data[j, n])
      }
    }
    z[[i]]= z.mat

    ### update theta
    for (j in 1:n.sample){
      draw= proposal.theta.i(theta[i-1, j])
      p.ratio.2= exp(posterior.theta.i(z[[i-1]][j,], draw, b= b, n.item, data[j,])-
                     posterior.theta.i(z[[i-1]][j,], theta[i-1, j], b= b, n.item, data[j,]))  # calculate the
ratio of posterior densities. The ratio was on log scale. The exponential function takes it back
to the regular scale of a ratio.
      tmp= runif(1, 0,1)
      if (tmp<p.ratio.2){ # to decide if a sample drawn at this iteration should be retained
        theta[i, j]= draw
      } else(
        theta[i, j]= theta[i-1, j]
      )
    }
  }
  return(list(pi= pi, z=z, b = b, theta=theta))
}
```

APPENDIX B

R code for Study II

```
# R code for a Metropolis-within-Gibbs algorithm for the MMR (item difficulties are
unknown)

library(gtools)


# create functions used to update pi.i (Gibbs sampler)
posterior.pi.i= function(z.i, n.group){
  # browser()
  alpha.new= alpha + table(factor(z.i, levels= 1:n.group))
  draw= rdirichlet(1, alpha.new)
  return(draw)
}


# create functions used to update z.ij (Gibbs sampler)
posterior.z.ij= function( pi.i, theta.i, b.j, x.ij){
  p.j= exp(theta.i- b.j)/ (1+ exp(theta.i- b.j))
  if (x.ij== 1){
    p.new= p.j*pi.i
    draw= rmultinom(1, 1, p.new)
    z.tmp= which(draw==1)
  }
  else if (x.ij==0){
    p.new.2= (1-p.j)*pi.i
    draw= rmultinom(1,1, p.new.2)
    z.tmp= which(draw==1)
  }
  return(z.tmp)
}


# create functions used to update theta.i (Metropolis step)
likelihood.theta.i= function(z.i, theta.i, b, n.item, x.i){
  # browser()
  l= rep(NA, n.item)
  for (i in 1:n.item){
    b.gj= b[,i][z.i[i]]
    p.jg= exp(theta.i- b.gj)/ (1+ exp(theta.i- b.gj))
    l[i]= dbinom(x.i[i], size=1, prob= p.jg, log= T)
  }
  return(sum(l))
```

```
}
prior.theta.i= function(theta.i){
  return(dnorm(theta.i, 0, 1, log= T))
}
posterior.theta.i= function(z.i, theta.i, b, n.item, x.i){
  post= likelihood.theta.i (z.i, theta.i, b, n.item, x.i)+ prior.theta.i(theta.i)
  return(post)
}
proposal.theta.i= function(theta.i){
  draw= rnorm(1, theta.i, sd= 1)
  return(draw)
}


# create functions used to update b.gj (Metropolis step)
likelihood.b.gj= function(x.j, theta, b.gj, z.j, g){ # for individuals who are in latent group g on
item j, calculate the log-likelihood of their responses to item j
  x.j.g= x.j[z.j==g]
  p.gj= exp(theta[z.j==g] - b.gj)/ (1+ exp(theta[z.j==g]- b.gj))
  l= dbinom(x.j.g, size=1, prob = p.gj, log = T)    # log-transformed
  return(sum(l))
}
prior.b.gj= function(b.gj){   # evaluate the prior density function of b.gj at a given sample of
b.gj
  return(dnorm(b.gj, 0, 1, log= T))
}
posterior.b.gj= function(x.j, theta, b.gj, z.j, g){ # calculate the posterior distribution on log
scale
  likelihood.b.gj(x.j, theta, b.gj, z.j, g)+ prior.b.gj(b.gj)
}

proposal.b.gj= function(b.gj){
  draw= rnorm(1, b.gj, sd= 1) # draw a sample from a proposal distribution
  return(draw)
}



# MCMC
run.mcmc = function(alpha, start.z, start.b, start.theta, n.group, n.sample, n.item, n.iter, data,
sum.difficulty){

  pi= vector('list', n.iter-1)
  z= vector('list', n.iter)
  b= vector('list', n.iter)    # create a list to store the samples of b obtained in all the iterations
  theta= matrix(NA, nrow= n.iter, ncol= n.sample)
```

```
  z[[1]]= start.z
  b[[1]]= start.b
  theta[1,]= start.theta

  for ( i in 2:n.iter){

    ### Update pi
    pi.mat= matrix(NA, nrow= n.sample, ncol=n.group)
    for (j in 1:n.sample){
      # browser()
      pi.mat[j,]= posterior.pi.i(z.i= z[[i-1]][j,], n.group)
    }
    pi[[i-1]]= pi.mat

    ### update z
    z.mat= matrix(NA, nrow= n.sample, ncol= n.item)
    for (j in 1:n.sample){
      for (n in 1:n.item){
        z.mat[j, n]= posterior.z.ij(pi[[i-1]][j,], theta[i-1, j], b[[i-1]][, n], data[j, n])
      }
    }
    z[[i]]= z.mat

    ### update theta
    for (j in 1:n.sample){
      draw= proposal.theta.i(theta[i-1, j])
      p.ratio.2= exp(posterior.theta.i(z[[i-1]][j,], draw, b= b[[i-1]], n.item, data[j,])-
                     posterior.theta.i(z[[i-1]][j,], theta[i-1, j], b= b[[i-1]], n.item, data[j,]))
      tmp= runif(1, 0,1)
      if (tmp<p.ratio.2){
        theta[i, j]= draw
      } else(
        theta[i, j]= theta[i-1, j]
      )
    }

    ### update b
    b.mat= matrix(NA, nrow= n.group, ncol= n.item) # In each iteration, samples are stored in a
matrix with rows as groups and columns as items.
    for (g in 1:n.group){
      for (t in 1:n.item){
        draw= proposal.b.gj(b[[i-1]][g, t])
        p.ratio.3= exp(posterior.b.gj(data[,t], theta[i-1, ], draw, z.j= z[[i-1]][, t], g)-
                       posterior.b.gj(data[,t], theta[i-1, ], b[[i-1]][g, t], z.j= z[[i-1]][,t], g)) # calculate
the ratio of the posterior densities. The ratio was on log scale. The exponential function takes it
back to the regular scale of a ratio.
```

```
        tmp= runif(1, 0, 1)
        if (tmp< p.ratio.3){ # to decide if a sample drawn at this iteration should be retained
          b.mat[g, t]= draw
        } else{
          b.mat[g, t]= b[[i-1]][g, t]
        }
       }
     }
     for (g in 1:n.group){ # at the end of each iteration, rescale the samples of b obtained in this
iteration
       constant = (sum.difficulty - sum(b.mat[g,]))/n.item
       b.mat[g,] = b.mat[g,] + constant
     }
     b[[i]]= b.mat
   }
   return(list(pi= pi, z=z, b=b, theta=theta))
}
```

APPENDIX C

R code for Study III

```
# R code for calculating D(δ)

mean.deviance= function(x, inter, burn, people, response.c, item){
  # x = an object returned by function run.mcmc()
  # iter = number of iterations in the MCMC chain
  # burn = number of iterations used as a burn-in
  # people = number of individuals
  # response.c = a binary response matrix with rows as individuals and columns as items
  log.l.mat= rep(NA, times= iter) # create a vector to store the sum of the log-likelihood of all
the responses given the samples of the parameters obtained at that iteration
  for (r in (iter-burn+1):iter) {
    mat.i= rep(NA, people)
    for (i in 1: people) {
        p= 1/(1+ exp(-(x[[4]][r, i] - x[[3]][[r]])))
        p.i.c= t(p)*response.c[i,]
        p.i.w= (1-t(p))*(1-response.c[i,])
        p.i= p.i.c+ p.i.w
        p.i.marginal= x[[1]][[r-1]][i, ] %*% t(p.i)
        mat.i[i]= sum(log(p.i.marginal))
    }
    log.l.mat[r] = sum(mat.i)
  }
  return(list(mean(log.l.mat, na.rm = T), var(log.l.mat, na.rm = T)))
}
```

```
# R code for calculating D(δ̄)

deviance.mean= function(x, people, response.c, item){
  # x stores point posterior estimates
  mat.i= rep(NA, people)
  for (i in 1: people) {
    p= 1/(1+ exp(-(x[[3]][i] - x[[2]])))
    p.i.c= p*response.c[i,]
    p.i.w= (1-p)*(1-response.c[i,])
    p.i= p.i.c+ p.i.w
    p.i.marginal= x[[1]][i, ] %*% t(p.i)
    mat.i[i]= sum(log(p.i.marginal))
  }
  return(sum(mat.i))
}
```