Robustness of Mixture IRT Models to Violations of Latent Normality

by

Sedat Sen

(Under the Direction of Allan S. Cohen and Seock-Ho Kim)

Abstract

Unlike the traditional item response theory (IRT) models, mixture IRT (MixIRT) models can be useful when subpopulations are suspected. The usual MixIRT model is typically estimated assuming normally distributed latent ability. Research on finite mixture models suggests that spurious latent classes can be extracted even in the absence of population heterogeneity if the distribution of the data is non-normal. In this study, we conducted two simulation studies and an empirical study to examine the robustness of MixIRT models to violations of latent normality. Single class IRT data sets were generated using different ability distributions and then analyzed with MixIRT models to determine the impact of these distributions on the extraction of latent classes. Results suggest that estimation of mixed Rasch models resulted in spurious latent class problems in the data when distributions were bimodal and uniform. Mixture 2PL and mixture 3PL IRT models were found to be more robust to latent non-normality. Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), were used to inform model selection. For most conditions, the performance of the BIC index was better than the AIC index for selection of the correct model.

Index words: Mixture item response theory, non-normal ability, spurious latent class

Robustness of Mixture IRT Models to Violations of Latent Normality

by

Sedat Sen

B.A., Hacettepe University, Turkey, 2006

M.A., The University of Georgia, 2010

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

2014

Robustness of Mixture IRT Models to Violations of Latent Normality

by

Sedat Sen

Approved:

Major Professors: Allan S. Cohen

Seock-Ho Kim

Committee: Zenqiu Lu

Gary J. Lautenschlager

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2014

Dedication

Annem ve babama...

Acknowledgments

I want to give a special thank to my family –my parents, my brothers, and my lovely sister– for their love and support. I always believe that there is nothing I cannot do with their incentives. They always made me feel a special person. They provided the motivations for me to pursue my degrees in the U.S. in the past years. Their love has made me what I am today.

I could not have completed this dissertation without the contributions of many. First, I would like to thank my advisors, Dr. Allan S. Cohen and Dr. Seock-Ho Kim, for mentoring me during my graduate career. I always felt very lucky to work with two great minds in psychometrics.

Dr. Kim was the person who made me learn and love item response theory (IRT) which is going to be my main research area in the rest of my career. He was always willing to take the time to explain any advanced topics in IRT with words I could understand. I never forget the long discussions we had in his office. His generous training and mentorship not only taught IRT and other measurement techniques but also taught me how to become a good scholar.

Dr. Cohen was the person who helped me to take the first steps in the mixture IRT research. He has been always with me whenever I lost myself in the complexity of Bayesian mixture IRT models. This dissertation could not have been written without his guidance and encouragement. I greatly appreciate his generous support, including research assistantships, travel funding, and permission to use the technological resources which made my study finish earlier.

Thank you to my committee members: Dr. Zenqiu Lu and Dr. Gary J. Lautenschlager. I appreciate their taking the time to serve on my committee with their helpful comments and suggestions.

Many thanks are due to Mustafa V. Nural, whose friendship and interest in helping me with computational difficulties I confronted during my simulation studies. I would also like to thank Gokhan Oztunc for being the one who understands my feelings and problems more than anyone while I was far away from my home. I would also like to thank Oztunc family for inviting me to great dinners in the past five years. I wish to thank my Turkish friends in Athens for making me feel like I am always at home.

Finally, I thank Almighty God (Allah c.c.) for providing me with such wonderful committee members, sincere friends, and a supportive family.

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Statement of the Problem

Item response theory (IRT) consists of a family of statistical models that attempt to describe the relationship between observed item responses and latent variables (Embretson & Reise, 2000). The first applications of IRT models were primarily in educational measurement (Lord & Novick, 1968), but they also have been applied extensively in several other areas such as psychology, public health, and personality assessment research. The standard, unidimensional IRT models are also known as a strong modeling methods because their successful applications require holding several assumptions such as unidimensionality, invariance, local independence, monotonicity, and existence of a continuous function that describes the relationship between the probability of correct response and latent trait (Reckase, 2009). For instance, the assumption that all examinees are drawn from a single homogenous population implies that one set of item characteristic curves (ICCs) can be used to describe the relationship between item responses and the underlying latent trait. When there might be subgroups of respondents with different response-trait relationships, however, it may be necessary to consider other modeling approaches. In cases in which assumptions are violated, more complex models may be needed to "more accurately reflect the complexity of the interactions between examinees and test items" (Reckase, 2009, p. 53). Mixture IRT (MixIRT; Rost, 1990; Mislevy & Verhelst, 1990) models, for example, may provide a more useful explanation for the relationship between test items and latent construct when the invariance assumption is violated (von Davier, Rost, & Carstensen, 2007).

MixIRT models can be useful when subpopulations are suspected, but which differ along some unmodeled latent variable. These models have been applied to study a number of psychometric issues such as detecting test speededness (Bolt, Cohen, & Wollack, 2002; Wollack, Cohen, & Wells, 2003) and differential item functioning (Cohen & Bolt, 2005; Cohen, Gregg, & Deng, 2005; Samuelsen, 2005), identifying different personality styles (von Davier & Rost, 1997), and identifying solution strategies (Mislevy & Verhelst, 1990; Rost & von Davier, 1993), as well as classifying response sets (Rost, Carstensen, & von Davier, 1997).

The MixIRT model is analytically based on finite mixture models. The popularity of finite mixture models has increased in educational and psychological sciences with extensions to several latent variable models including latent class analysis (LCA; Clogg, 1995), structural equation models (Arminger, Stein, & Wittenberg, 1999; Jedidi, Jagbal, & DeSarbo, 1997), growth mixture models (Li, Duncan, & Duncan, 2001), factor mixture analysis (FMA; Lubke & Muthén, 2005) and MixIRT models (Rost, 1990; Mislevy & Verhelst, 1990). A comprehensive taxonomy of the different types of latent variable mixture models are provided by Muthén (2008). A general use of mixture models is to explain the underlying heterogeneity in the data by allocating this heterogeneity to two or more latent classes. One issue that has arisen with the use of these models, however, is that the extracted classes may not always reflect a heterogeneous population structure (Bauer & Curran, 2003). Rather, it may reflect some extraneous characteristic of the data. For instance, research on finite mixture models suggests that non-normal distributions in the data may produce spurious latent classes even in the absence of population heterogeneity (McLachlan & Peel, 2000; Bauer & Curran, 2003). In the case of non-normality, in other words, it has been shown that it is possible to obtain spurious latent classes. Similarly, violations of other model specific assumptions in these mixture models may also result in spurious classes (Alexeev, Templin, & Cohen, 2011).

When the number of classes within a population is not known a priori, the usual practice is to conduct an exploratory analysis to determine what the number of latent classes may be.

This exploratory approach is done by fitting models with an increasing number of classes to the data and then finding the best fitting among these candidate models. Selecting the best fitting model among the alternatives can be done using information-based indices such as Akaike's (1974) information criterion (AIC), Schwarz's (1978) Bayesian information criterion (BIC) or a hypothesis test based on likelihood ratio test statistic (McLachlan & Peel, 2000). The optimum model for the data is determined based on comparison of fit indices among those models being considered.

The issue of extracting the correct number of classes has become a longstanding and unresolved issue for researchers who apply finite mixture models in their research. Given that the use of mixture models has increased in the educational and psychological research, it is important to ensure the detection of the correct number of latent classes in the data. Most of the research on over extraction within the latent variable modeling context has focused on either the violation of model specific assumptions (Alexeev et al., 2011; Bauer, 2007; Bauer & Curran, 2003) or model selection statistics (Li, Cohen, Kim, & Cho, 2009; Nylund, Asparouhov, & Muthén, 2007; Tofighi & Enders, 2007; Wall, Guo, & Amemiya, 2012). These studies also differed in the types of models inspected.

Bauer and Curran (2003) conducted a simulation study to examine the detection of spurious latent classes appear in growth mixture models (GMM). This was done by analyzing single-class data sets generated with normal and non-normal distributions. The results of that research showed that the one-class solution had better fit for the data drawn from normal distribution but the two-class solution had better fit for the data drawn from non-normal distributions. Bauer and Curran concluded that the over extraction was observed because a variety of non-normal distributions can be approximated by a mixture of normal distributions, even in the absence of true population subgroups. Bauer (2007) focused on the core assumptions of the GMM including within-class conditional normality, correct specification of covariance structure to accurately reproduce the means, variances, and covariances of the

repeated measures, within-class linear relationship between individual trajectory parameters and exogenous predictors, missing at random assumption for missing data, and independence of sampled individuals. Bauer examined the sensitivity of the LCA model when these assumptions are not met, particularly the effect of these violations on the number of estimated latent classes. Bauer noted that the impact of non-normality is potentially of importance with test data as data from most tests do not exhibit conditional normality. As a result, even mild violations of conditional normality resulted in the estimation of spurious latent classes. It seems apparent, therefore, that the number of latent classes extracted from the data may represent either true distinct subgroups or simply a result due to the mixture model attempting to accommodate the non-normality in the data.

Tofighi and Enders (2007) provide a comprehensive evaluation of nine different fit indices (information criteria and likelihood based statistics) with a simulation study within the context of GMMs. Five factors were manipulated in the study: the number of time points in the data, sample sizes, separation of the latent classes, the mixing percentage, and within class distribution shape. Results from that study indicated that the sample-size adjusted BIC (SABIC; Sclove, 1987) and the Lo-Mendell-Rubin (LMR; Lo, Mendell, & Rubin, 2001) likelihood ratio test are promising in determining the number of classes.

Nylund et al. (2007) conducted a similar Monte Carlo simulation study to compare the performance of information criteria and hypothesis tests based on likelihood ratio test statistic used for determining the number of classes in mixture modeling. These indices were compared on solutions for three different mixture model analyses, a LCA, factor mixture models (FMMs), and GMMs. Results showed that for the LCA models with continuous outcomes, FMM and GMM models, the bootstrap likelihood ratio test (BLRT) performed better than LMR and likelihood-ratio tests for determining the correct number of classes. Results also indicated that BIC was superior to AIC and consistent AIC (CAIC; Bozdogan, 1987) for all three types of mixture model analyses. Li et al. (2009) compared five fit indices

on Bayesian estimates of dichotomous mixture Rasch, 2-parameter and 3-parameter models. Results from Li et al. indicated that BIC performed better in most cases. Other indices compared were deviance information criterion (Spiegelhalter, Best, & Carlin, 1998), AIC, pseudo Bayes factor (PsBF), and posterior predictive model checks (PPMC).

Alexeev et al. (2011) also examined the effects of violation of model assumptions on extracting the correct number of class in MixIRT models. Data were generated with a two parameter logistic (2PL) IRT model. When estimated with a mixture Rasch model (MRM), results indicated that a two-class solution was a better fit than a one-class 2PL model. In addition, items that did not conform to assumptions of the Rasch model (RM; Rasch, 1960) also were found to be a source of spurious latent classes. Specifically, when a MRM was applied to data generated with a one-class 2PL, one or more additional latent classes were detected. Alexeev et al. found that a single item with high discrimination was sufficient to spawn a second class when the data were analyzed with a MRM.

As reported in Bauer and Curran (2003), Bauer (2007), and Alexeev et al. (2011), even small departures from model assumptions may have effect on the number of latent classes identified as well as on model parameter estimates. Although Alexeev et al. (2011) demonstrated that the violations of RM assumption (i.e., equal discrimination) produced spurious classes, there has been no study conducted to examine the effect of different distributions of the latent variable on the number of classes identified in MixIRT models. As Bauer and Curran (2003) showed with the GMM, it is important to know whether the non-normality was responsible for generating additional latent classes from the data in MixIRT models.

Similar to the normality assumption made in GMM analysis, the usual MixIRT model is also typically estimated with maximum likelihood (ML) based methods which assume within class normality. Research on finite mixture models suggests that latent classes potentially can be extracted, even when there is no population heterogeneity, if the distribution of the data is non-normal (Bauer & Curran, 2003). Empirical evidence suggests, in fact, that ability

distributions may not always be normal in educational applications of IRT to populations with specific characteristics (Seong, 1990).

The assumption of normality is made for a number of statistical models for either dependent variables or residuals (Cohen, Cohen, West, & Aiken, 2002). Latent variable modeling applications also rely on the estimation methods that require normality assumption. Typically, IRT models are estimated using marginal maximum likelihood (MML) estimation based on some known latent trait distribution (Bock & Aitkin, 1981; Bock & Lieberman, 1970). ML estimates of item and person parameters are typically estimated assuming a normal ability distribution when MML estimation is used rather than conditional maximum likelihood (CML). A standard normal prior is often assumed on the ability parameters when using Bayesian estimation methods. However, researchers commonly use the ML method or normal prior without checking the distributional assumption. For instance, research has shown that the latent non-normality assumption may be more reasonable in psychopathology data in which the level of psychopathology is low for most individuals, medium for some individuals and high for only a few individuals (Woods & Thissen, 2006, p. 282). IRT models provide best estimates when the true distribution of the data matches with the assumed distribution (Woods, 2004). Parameter estimates will be biased, in other words, when IRT models estimated with normality based methods are fit to data where the latent distribution is non-normal (Bock & Aitkin, 1981; Zwinderman & van den Wollenberg, 1990). Furthermore, evidence suggests that the normality assumption for latent trait might not be optimal for best parameter recovery (Stocking, 1990; Wingersky & Lord, 1984). It has been showed that a rectangular distribution of abilities yields reduced standard errors for item parameters than a normal ability distribution (Wingersky & Lord, 1984).

It is not possible to know the underlying distribution of ability parameters, because it is not directly observable, however, the effect of a non-normal ability distribution on estimates of the parameters can be observed. Follman (1988) demonstrated that the nonparametric marginal logistic model with no parametric assumptions on ability distribution fit data better than the model with normal distribution assumption. As cited in Woods (2004), same data set was analyzed by Thissen (1991) and found to have a skewed latent ability distribution. Since the distribution of ability parameters is unobservable, the latent non-normality situation was mostly studied with simulation analysis in the IRT literature. A list of these studies can be found in Woods (2004).

Although the latent ability distribution is unobservable, the observed test score distributions may give information about the latent ability distribution. Sinharay, Johnson, and Stern (2006) and Sinharay (2005) show that raw score distributions are related to true ability distribution and can be used for detecting misfit of an IRT model. Raw score distributions of test scores can be normal, platykurtic, skewed, or U shaped depending on the purpose of the test, however, it is more convenient to assume normality in most of the statistical analyses. In reality, the normality assumption will often be incorrect "when individuals are non-randomly sampled from the population (e.g., scores are obtained only from students in lower-level classes) or when the population distribution itself is non-normal (e.g., scores are obtained from extremely easy or extremely difficult tests)" (Sass, Schmitt, & Walker, 2008, p. 66). Measurement properties, such as floor or ceiling effects, may also result in non-normality (Bauer & Curran, 2004). As Nunnally (1978) stated in his psychometric theory book,"test scores are seldom normally distributed, even if the number of items is large" (p. 160). Similarly, Micceri (1989) demonstrated that the presence of non-normal distributions of ability is more realistic for achievement tests because the shapes of ability distribution in most of the achievement tests are more likely to be skewed. A number of studies have shown that skewed, platykurtic, uniform and multimodal distributions are common distributions

observed in educational data. For instance, Gregoire and Driver (1987) investigated several populations and found that almost 40 percent of the distributions exhibited skewness greater than 0.94. Positively skewed distributions can appear in the data when examinees are mostly low-to-moderate ability, whereas negatively skewed data can be observed when examinees are mostly moderate-to-high ability (Nandakumar & Yu, 1996). Typically, non-normal data appear to have skewness less than 0.8 and kurtosis between -0.6 and 0.6 based on empirical observations (Fleishman, 1978; Pearson & Please, 1975). Platykurtic data are observed when the test score distribution shows a high negative kurtosis values. As mentioned earlier, it is also possible to observe bimodal distributions when two radically different groups of examinees are administered the same test (Nandakumar & Yu, 1996) or when items are designed to discriminate sharply between masters and nonmasters (Subkoviak, 1976). Further, multimodal distributions can be observed when different grade levels are present (Subkoviak, 1976). As another example, bimodal distributions can appear in achievement test scores from both a privileged and an underprivileged school (Sass, Schmitt, & Walker, 2008). Additionally, Micceri (1989) showed that almost one third of the distributions studied were either bimodal or multimodal based on reviews of a total of 440 distributions from several test centers and journal articles published between the years 1982 and 1984. Nearly all of the distributions were significantly non-normally distributed.

1.2 Purpose of the Study

Given that a significant proportion of test score distributions are likely to be non-normal, it is important to examine the effect of potential non-normality on the extraction of latent classes by MixIRT models. The impact of distributional conditions on the extraction of latent classes was examined in MixIRT models. In particular, the sensitivity of Markov chain Monte Carlo (MCMC) estimation in the dichotomous MixIRT models to latent non-normality was examined. Item response data were generated by a given single class IRT model with different

ability distributions and then were analyzed with mixture Rasch, mixture 2-parameter, and mixture 3-parameter logistic models using normality based methods to determine the impact of departure from normality on the extraction of latent classes.

Chapter 2

THEORETICAL FRAMEWORK

This chapter begins with a brief description of finite mixture models. Parameterizations of MixIRT models used in this study are presented next. Additionally, methodological issues, estimation and model selection in MixIRT models are described. This chapter ends with a section on latent non-normality issue in IRT models.

2.1 Finite Mixture Models

Mixture models have been extensively used in several areas including, biology, economics, psychology and education since the contribution of Newcomb (1886) and Pearson (1894). The general purpose of mixture models is to determine the number of underlying groups and to estimate their proportions and specific parameters. The basic definition of finite mixture models is presented in several textbooks (Everitt & Hand, 1981; McLachlan & Peel, 2000; Titterington, Smith, & Makov, 1985). Following the definition in McLachlan and Peel (2000; pp. 6–7), we assume that $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ denote a random sample of size n, where $\boldsymbol{Y}_j$ is a p-dimensional random vector with probability density function $f(\boldsymbol{y}_j)$ on $\mathbb{R}^p$. The density for a g component finite mixture model is defined as

$$f(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i), \tag{2.1}$$

where $f_1(y_j), ..., f_g(y_j)$ are the component densities of the mixture, parameters $\pi_1, ..., \pi_g$ are often called nonnegative mixing proportions with $0 \leq \pi_i \leq 1$ for $(i = 1, ..., g)$ and $\sum_{i=1}^{g} \pi_i = 1$ and where $\boldsymbol{\Psi}$ is the vector of all the parameters in the mixture model,

$$\boldsymbol{\Psi} = (\pi_1, \pi_2, ..., \pi_g, \boldsymbol{\Theta}')'. \tag{2.2}$$

Additionally, $\boldsymbol{\Theta}$ is a vector containing the $\theta_1, ..., \theta_g$ model parameters and $\boldsymbol{\Theta}' = (\boldsymbol{\theta}_{1g}, ..., \boldsymbol{\theta}_{ng})'$ represents a vector of unknown parameters for the gth latent class. Typically, neither the conditional densities $f(\boldsymbol{y}_j)$ nor the mixing proportions π_i are known. McLachlan and Peel (2000) provide a way of generating a random vector of $\boldsymbol{Y}_j$ with a g-component mixture density $f(\boldsymbol{y}_j)$. Let $\boldsymbol{Z}_j$ be a categorical random variable taking on the values $1, ..., g$ with the probabilities $\pi_1, ..., \pi_g$. Suppose the conditional density of $\boldsymbol{Y}_j$ given $\boldsymbol{Z}_j = i$ is $f_i(\boldsymbol{y}_j)(i = 1, ..., g)$. If the i-th observation of $\boldsymbol{Y}_j$ comes from the g-th class, then $z_{ij} = 1$, otherwise $z_{ij} = 0$. Then, $\boldsymbol{Z}_j$ has a multinomial distribution and one draw on g categories with probabilities $\pi_1, ..., \pi_g$ can be parameterized as follows:

$$Prob(\boldsymbol{Z}_j = \boldsymbol{z}_j) = \pi_1^{z_{1j}}, \pi_2^{z_{2j}}, ..., \pi_g^{z_{gj}}. \tag{2.3}$$

Bauer and Curran (2003) note two purposes of finite mixture models: (i) the mixture models can identify distinct subgroups in the population or (ii) they may approximate complex distributions with simpler component distributions. Parameter estimations of finite mixture models can be done using maximum-likelihood estimation (MLE) and Bayesian estimation.

The latent class model (LCM; Lazarsfeld & Henry, 1968) is a special case of finite mixture models used for categorical response data. The LCM is used to identify latent groups that differ qualitatively. Under the assumption of the existence of G latent classes, the marginal probability of a response pattern $(x_1, ..., x_I)$ is given as

$$P(x_1, ..., x_I) = \sum_{g=1}^{G} \pi_g \prod_{i=1}^{I} p_{gi}(x_i), \tag{2.4}$$

where π_g is the probability that an individual is a member of class g, as noted above, this is also known as the mixing proportion. Under the assumption of local independence of

response variables the marginal probability of a response pattern can be written as follows:

$$P(x_1, ..., x_I|g) = \prod_{i=1}^{I} p_{gi}(x_i), \tag{2.5}$$

where p_{gi} is the probability of a correct response to item i from an examinee in class g. LCM allows us to assign a posterior probability for each individual. Individuals are assigned to the most likely class based on the posterior distribution. Using Equation 2.4 and applying Bayes' rule we can write the posterior probability of class membership as follows:

$$P(g^{'}|x_1, ..., x_I) = \frac{\pi(g^{'})p(x_1, ..., x_I|g^{'})}{\sum_{g=1}^{G} \pi(g)p(x_1, ..., x_I|g)}, \tag{2.6}$$

for latent class $g^{'}$. The unconstrained LCM model equation can be re-parameterized using a logistic transformation approach. Under this re-parameterization, the conditional probabilities can be written as

$$p_{gi}(x) = \frac{\exp(\beta_{gix})}{1 + \sum_{y=1}^{M_i} \exp(\beta_{giy})} = \exp(\beta_{gix} - \zeta_{ci}), \tag{2.7}$$

where

$$\zeta_{ci} = \ln[1 + \sum_{y=1}^{M_i} \exp(\beta_{giy})]. \tag{2.8}$$

This parameterization allows us to make additional constraints (von Davier & Rost, 2007). Similarities between LCM and IRT can be seen easily from this logistic form. MixIRT models can be obtained as a constrained LCM. In the formula above, β_{gix} includes both item and person effects. As a result, as the number of classes increases, the number of parameters also increases. Similar to IRT parameterization, a more parsimonious model can be obtained if item and class parameters are written instead of β_{ixg}

$$\beta_{ixg} = \theta_g - \alpha_{ix}, \tag{2.9}$$

similarly

$$\beta_{ixg} = x(b_i\theta_g - a_i), \tag{2.10}$$

which will be similar to a 2PL IRT model (von Davier & Rost, 2007). In the following sections, MixIRT models are presented using this form in the conditional probability (see Equation 2.7).

2.2 Mixed Rasch Model

In order to overcome the deficiencies associated with traditional dichotomous IRT models, Rost (1990) developed a model called mixed Rasch model (MRM) that combines a RM with an LCM. Both the RM and LCM assume local independence; however the latent variable is continuous in the former whereas it is categorical in the latter. Rasch (1960) defined the probability of a correct response as follows:

$$P(u_{ij} = 1|\xi_{ij}) = \frac{\xi_{ij}}{1 + \xi_{ij}}, \tag{2.11}$$

where u_{ij} is the dichotomous item response scored as 1 for correct response and 0 for incorrect, and ξ_{ij} is the ratio of ability, η_j, and item difficulty, δ_i. Similarly, using the following equalities $\eta_j = e^{\theta_j}$ and $\delta_i = e^{\beta_i}$ with logarithmic transformations (Baker & Kim, 2004), one-parameter logistic RM for dichotomous type items can be written as

$$P_i(\theta) = \frac{e^{(\theta_j - \beta_i)}}{1 + e^{(\theta_j - \beta_i)}}, \tag{2.12}$$

where θ_j is the ability parameter for examinee j which is assumed to drawn from a single population and β_i is the item difficulty parameter for item i. This model assumes unidimensionality and local independence of items.

Based on the MRM, a unidimensional RM is assumed for each latent class. However, each class has different set of item and ability parameters. As a result, this model relaxes the invariance property of item parameters (von Davier, Rost, & Carstensen, 2007). The basic assumption in the MRM is that the observed item response data come from a composite population that can be subdivided into mutually exclusive and exhaustive latent classes (Rost, 1990; von Davier & Rost, 2007). Unlike the non-mixture (i.e., single group) RMs, the classes in a MRM reflect both quantitative and qualitative differences. The quantitative differences occur along the latent ability modeled by the RM in each latent class and with the same interpretation. The qualitative variable is a categorical variable which determines latent class membership. Thus, this model includes both the RM parameters and LCM parameters. Similar to the LCM equation, the probability of a given response pattern can be defined as

$$p(x_1, ..., x_I|\theta_j) = \sum_{g=1}^{G} \pi_g p(x_1, ..., x_I|\theta, g). \tag{2.13}$$

Under the assumption of local independence, the marginal probability can be written as:

$$p(x_1, ..., x_I|\theta_j, g) = \prod_{i=1}^{I} \frac{\exp[x_i(\theta_{jg} - \beta_{ig})]}{1 + \exp[x_i(\theta_{jg} - \beta_{ig})]}. \tag{2.14}$$

Each examinee is characterized by a latent class parameter, g, and an ability parameter, θ_{jg}, and similarly, each item is characterized by a latent class parameter g and a difficulty parameter, β_{ig}. More specifically, the conditional probability of a correct response in the MRM is given as

$$P(x_{ij} = 1|\theta_j) = P_{ij} = \sum_{g=1}^{G} \pi_g \frac{\exp(\theta_{jg} - \beta_{ig})}{1 + \exp(\theta_{jg} - \beta_{ig})}, \tag{2.15}$$

where x_{ij} is the 0/1 response of examinee j to item i (0 = incorrect response, 1 = correct response), π_g is the proportion of examinees for each class, θ_{jg} denotes ability of examinee j within latent class g, and β_{ig} denotes difficulty of item i within latent class g.

In order to ensure identification, certain constrains on item difficulty parameters and mixing proportions should be made. For the MRM, Rost (1990) has proposed that $\sum_i \beta_{ig} = 0$. Further, since the model is a discrete mixture model, $\sum_g^G \pi_g = 1$ with $0 < \pi_g < 1$. The probability of belonging to each of the latent classes can also be estimated for every individual using the posterior probability as in Equation 2.6. Individuals are then assigned to the latent class with the highest probability.

It is possible that in many testing situations individual persons or subgroups of persons may behave differently than the majority of persons in the population. In such a case, it may be less plausible to use typical IRT models. An alternative is then to use a MixIRT model (e.g., MRM) allowing differences between the parameters within each latent class. These models are very useful in detecting qualitative differences in the response patterns. The combination of an IRT model with a LCM allows each latent class to have the same IRT model with different parameters (Rost, 1990). The MRM differs from the unidimensional RM model in that the item and ability parameters are conditional on latent class g. Each item has the same parameter across all examinees in the RM. Thus, the MRM permits each latent class to have different set of item parameters.

MRMs provide a simultaneous estimation of a latent ability and latent class membership for all examinees. The MRMs have been applied in a number of studies (Mislevy & Verhelst, 1990; Rost, Carstensen, & von Davier, 1997; Rost & von Davier, 1993; von Davier & Rost, 1997; Yamamoto & Everson, 1997). Polytomous extensions of MRM have also been developed (e.g., Eid & Zickar, 2007; von Davier & Rost, 1995; von Davier & Yamamoto, 2004).

2.3 Mixture IRT Models

The 3PL IRT model was proposed by Birnbaum (1968) to accommodate guessing. It does this by adding a lower-asymptote parameter, denoted as γ. The probability of a correct response for a 3PL model can be described as

$$P(x_{ij}=1|\theta_j)=P_{ij}=\gamma_i+(1-\gamma_i)\frac{\exp[\alpha_i(\theta_j-\beta_i)]}{1+\exp[\alpha_i(\theta_j-\beta_i)]}. \tag{2.16}$$

The 2PL IRT model can be seen as a constrained case of 3PL model in which it is assumed that the lower-asymptote parameter is zero (i.e., no guessing). The probability of a correct response for a 2PL IRT model can be defined as:

$$P(x_{ij}=1|\theta_j)=P_{ij}=\frac{\exp[\alpha_i(\theta_j-\beta_i)]}{1+\exp[\alpha_i(\theta_j-\beta_i)]}. \tag{2.17}$$

Both of these models relax the equal discrimination assumption of the 1PL model and the assumption that the discrimination is 1.0 of the RM. Extensions of these models to mixture distribution IRT models are very straightforward. The probability of a correct response in a mixture 2PL (Mix2PL) IRT model can be written as

$$P(x_{ij}=1|\theta_j)=P_{ij}=\sum_{g=1}^{G}\pi_g\frac{\exp[\alpha_{ig}(\theta_{jg}-\beta_{ig})]}{1+\exp[\alpha_{ig}(\theta_{jg}-\beta_{ig})]}. \tag{2.18}$$

where α_{ig} denotes the discrimination of item i in class g. As noted above, MixIRT models permit different item parameters in each latent class. For the Mix2PL model, this means that both the item difficulty and item discrimination parameters are permitted to be class-specific.

Similarly, the mixture 3PL (Mix3PL) IRT model is assumed to describe unique response propensities for each latent class. This model also allows item guessing parameters to differ in addition to item difficulty and discrimination parameters. As for the MRM and Mix2PL model, each latent class also can have different ability parameters. The probability of a correct response for a Mix3PL model can be described as

$$P(x_{ij}=1|\theta_j)=P_{ij}=\sum_{g=1}^{G}\pi_g\left(\gamma_{ig}+(1-\gamma_{ig})\frac{\exp[\alpha_{ig}(\theta_{jg}-\beta_{ig})]}{1+\exp[\alpha_{ig}(\theta_{jg}-\beta_{ig})]}\right). \tag{2.19}$$

where π_g are the mixing proportions, θ_{jg} is the class-specific ability parameter for person j in class g and γ_{ig} is guessing parameter for item i in class g. The MixIRT models have been applied in a number of studies (Cohen & Bolt, 2005; Li et al., 2009).

2.4 Estimation Methods for MixIRT Models

Two estimation algorithms are described in this section, a ML algorithm and a Bayesian algorithm. In traditional IRT models, parameter estimation can be done using MLE techniques such as joint maximum likelihood estimation (JMLE), marginal maximum likelihood estimation (MMLE; Bock & Aitkin, 1981; Bock & Lieberman, 1970) and conditional maximum likelihood estimation (CMLE). In principle, ML estimates of the item parameters and the ability parameters can be obtained using standard numerical optimization procedures (e.g., Newton-Raphson). These procedures are based on iterations which aim to find the parameter estimates that maximize the likelihood function. JMLE estimates the structural (item) parameters and nuisance (person) parameters together; CMLE and MMLE can estimate item parameters without estimating person parameters. Although CMLE is limited to the family of Rasch models, MMLE can be used for all type of models within IRT framework. Bock and Aitkin (1981) proposed an algorithm called expectation maximization (EM; Dempster, Laird, & Rubin, 1977) which can be used to obtain CML and MML estimates of parameters in the IRT models (von Davier & Rost, 2007). MLE methods are discussed below. In addition, Bayesian solutions are also available for IRT parameter estimations. These are discussed below as well.

Parameter estimation of MixIRT models can be more complex than the traditional IRT models due to the increase in the number of parameters that need to be estimated. Parameter estimation for MixIRT models can also be applied either using MLE or MCMC methods in the Bayesian context. Unlike MLE, parameter estimation with MCMC methods relies on a priori distributional assumptions on person and item parameters that allow us to update

beliefs about parameters. MCMC has been found to be useful for estimation of complex MixIRT models, because it does not require the integration of the likelihood function, which can be difficult when many parameters are estimated (Junker, 1999). Another advantage of Bayesian estimation is that the local maxima problem observed in MLE does not arise with MCMC (Dai, 2009). On the other hand, the MCMC approach is often time consuming to implement due to computational intensity especially for MixIRT models (Li et al., 2009). MLE does not typically require the time that MCMC does and has also been used successfully in estimating some MixIRT models (von Davier & Rost, 2007). Unbounded likelihood functions and sensitivity to poor starting values also can be problematic in MLE. These problems can cause inconsistent estimates (Cheng & Traylor, 1995). On the other hand, typical issues such as improper posterior distributions due to improper priors might occur in the MCMC estimation (Cho, Cohen, & Kim, 2013; McLachlan & Peel, 2000).

Fortunately, there are several estimation methods and software packages available for estimating some MixIRT models. These include the JMLE method as implemented in the R package mixRasch (Willse, 2009), and the MMLE method as implemented along with the EM algorithm for the MRM and Mix2PL models in the computer software Mplus (Muthén & Muthén, 2011). The computer software mdltm (von Davier, 2005) also implements the EM algorithm with MMLE as does the computer software Latent GOLD (Vermunt & Magidson, 2005). The CMLE method is implemented in the Winmira software (von Davier, 2001), the mRm (Preinerstorfer, 2011) and the psychomix (Frick, Strobl, Leisch, & Zeileis, 2012) packages in R software. MCMC estimation is possible using the WinBUGS computer software (Spiegelhalter, Thomas, & Best, 2003), Mplus and proc MCMC in SAS software (SAS Institute, 2008).

All of these packages allow for estimation of MRM. The Mix2PL IRT model can be fit using Latent GOLD, Mplus and WinBUGS programs, however, only WinBUGS program has the capability at this time of estimating Mix3PL IRT model. Thus, the computer software

WinBUGS was used in this study for estimating all the models to be studied. Although only MCMC estimation method was used in this study, a general description of MLE methods is presented below. Following that, MCMC estimation methods will be described.

2.4.1 MLE Methods

Although, CMLE and MMLE methods are commonly used for MixIRT models, JMLE method can also be applied in MixIRT models. Willse (2011) demonstrated that MRMs can be analyzed using an R package called mixRasch (Willse, 2009). Currently, this is the only software that can be used to estimate the MRM using JMLE. The EM algorithm is used to estimate model parameters.

CML estimates of MRMs can be estimated using EM algorithm (Rost, 1990, 1991; Rost & von Davier, 1995; von Davier & Rost, 1995). Item parameters of each class are estimated within each Maximization step (M-step) and expected pattern frequencies are estimated in the Expectation step (E-step) for each class using the observed pattern frequencies and the estimates from M-step (Rost, 1990). This process is outlined in Rost (1990) and Rost and von Davier (1995). Following their description, let the conditional probability of a correct response to a RM be

$$P(\boldsymbol{X} = \boldsymbol{x}|g) = \int_{-\infty}^{\infty} \prod_{i=1}^{k} \frac{\exp[x_i(\theta_g + \beta_{ig})]}{1 + \exp(\theta_g + \beta_{ig})} dF_g(\theta_g) \tag{2.20}$$

and

$$P(\boldsymbol{X} = \boldsymbol{x}) = \sum_{g=1}^{G} \pi_g P(\boldsymbol{X} = \boldsymbol{x}|g). \tag{2.21}$$

Using the elementary symmetric functions z_{rg} of order r in class g, captures all possible response patterns to obtain a certain score, and the continuous variables θ_g are conditioned out:

$$P(\boldsymbol{X} = \boldsymbol{x}|g) = \pi_{r|g} \frac{\exp(\sum_{i=1}^{k} x_i \beta_{ig})}{z_r[\exp(\boldsymbol{\beta}_g)]}. \tag{2.22}$$

As can be seen, there are three parameters to be estimated in this formula; π_g, $\pi_{r|g}$, and β_{ig}. These parameters are estimated in the M-step by maximizing the log-likelihood function:

$$\log L_g = \sum_x \hat{n}(\boldsymbol{x}|g)\{\log \pi_{r|g} + \sum_{i=1}^{k} x_i \beta_{ig} - \log[z_r(\exp \beta_g)]\}. \tag{2.23}$$

When setting the first partial derivatives to zero we can obtain following estimation equations for item parameters:

$$\hat{\beta}_{ig} = \log \frac{n_{ig}}{\sum_{r=0}^{k} n_r z_{r-1}/z_r}. \tag{2.24}$$

This equation can be used to iteratively improve item parameter estimates. Additionally, assuming item parameters are known from the first step, the population parameters, class size and class probabilities, can be estimated using $\hat{\pi}_g = n_g/N$ and $\hat{\pi}_{r|g} = n_r/n_g$, respectively where n_r is the observed sample frequency of score r and N is the total sample size (see also von Davier & Rost, 1995). CMLE does not assume any population distribution by conditioning out the ability parameter (von Davier & Yamamoto, 2007). As a result, the number of parameters does not increase with sample size as it does for JMLE.

Mislevy and Verhelst (1990) demonstrated the use of MMLE for a MRM. As with CMLE, parameter estimates can be obtained through use of the EM algorithm. However, the MMLE method assumes a distribution (e.g., normal) on person ability (θ) and integrates the likelihood over the assumed ability distribution. MMLE is based on the maximization of the marginalized likelihood function. The probability of response pattern x_{ij} conditional on person parameters (i.e., ζ_{jg} and θ_{jg}) can be written as

$$p(x_{ij}|\zeta_{jg}, \theta_j, \alpha) = \prod_g \left\{ \prod_i [f_g(\theta_{jg}, \beta_{ig})]^{x_{ij}} [1 - f_g(\theta_{jg}, \beta_{ig})]^{1-x_{ij}} \right\}^{\zeta_{jg}}, \tag{2.25}$$

where x_{ij} is the response vector coded as 1 (correct) or 0 (incorrect), β_{ig} is the difficulty parameter for item i within class g and θ_{jg} is the ability parameter for person j for class g. ζ_{jg} equals one if person j appears in class g and zero otherwise. The marginal probability of

response vector $\mathbf{X}$ can be written as:

$$p(\boldsymbol{X}|\alpha,\pi,\eta) = \sum_g \pi_g \int p(\boldsymbol{X}|\theta_g,\zeta_g = 1,\alpha) g_g(\theta_g|\eta_g) d\theta_g, \tag{2.26}$$

where α, π, and α are structural parameters. The likelihood for structural parameters given a response vector $\mathbf{X}$ of N persons can be shown as the product of marginal probabilities under the assumption of local independence. Thus, the marginalized loglikelihood function can be defined as:

$$LL = \sum_j \log \sum_g \pi_g \int p(\boldsymbol{X}|\theta_g,\zeta_g = 1,\alpha) g_g(\theta_g|\eta_g) d\theta_g. \tag{2.27}$$

As in traditional IRT models, a Newton-Raphson iterative process is used to obtain parameter estimates in addition to the EM algorithm (see also Mislevy & Verhelst, 1990).

2.4.2 MCMC Estimation

MCMC estimation uses a Markov chain to simulate observations in order to make inferences about the parameters (Patz & Junker, 1999). This algorithm tries to obtain a stationary posterior distribution where the chain converges. Each stage in a Markov chain represents a sample from the joint posterior distribution of the parameters. Final estimates of parameters can be taken as either the mean or the mode of the posterior distribution. Posterior distributions of parameter estimates are simulated iteratively using sampling algorithms such as Gibbs sampling, adaptive rejection sampling (ARS; Gilks & Wild, 1992), and the Metropolis-Hastings algorithm within Gibbs (M-H; Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). A Gibbs sampling algorithm samples parameter from its conditional distribution when other parameters are fixed (Albert, 1992). This approach is very useful when the full conditional distributions are familiar forms (e.g., normal). However, the ARS or M-H algorithms may be preferred over Gibbs sampling when conditional distributions cannot be obtained in known form (Patz & Junker, 1999). This is often observed in IRT

models because posterior distributions of IRT models are determined up to a normalizing condition resulting in log-concave full conditional distributions (Bolt, Cohen, & Wollack, 2001).

The MCMC algorithm is a Bayesian estimation algorithm. This algorithm proceeds as follows: Let $\mathbf{X}$ be observed data and θ be model parameters. Then, the joint density of $\mathbf{X}$ and θ can be defined as:

$$P(\boldsymbol{X}, \theta) = P(\boldsymbol{X}|\theta)P(\theta), \tag{2.28}$$

where $P(\theta)$ denotes a prior density for θ and $P(\boldsymbol{X}|\theta)$ is the likelihood function. We can produce the following equality using Bayes' law:

$$P(\theta)P(\boldsymbol{X}|\theta) = P(\boldsymbol{X})P(\theta|\boldsymbol{X}). \tag{2.29}$$

This equation can be rearranged as follows:

$$P(\theta|\boldsymbol{X}) = \frac{P(\boldsymbol{X})P(\boldsymbol{X}|\theta)}{P(\boldsymbol{X})}. \tag{2.30}$$

We can obtain the posterior distribution by applying Bayes' theorem as follows:

$$P(\theta|\boldsymbol{X}) = \frac{P(\theta)L(\theta|\boldsymbol{X})}{\int P(\theta)L(\theta|\boldsymbol{X})d\theta}, \tag{2.31}$$

where $L(\theta|\boldsymbol{X})$ is the likelihood function and $\int P(\theta)L(\theta|\boldsymbol{X})d\theta$ is used instead of $P(\boldsymbol{X})$ by integrating the numerator over θ. This expression is called as normalizing constant that ensures that $P(\theta|\boldsymbol{X})$ integrates to one (Gill, 2002). In other words, the posterior distribution is proportional to the product of the likelihood and the prior distribution:

$$\textit{Posterior distribution} \propto \textit{Likelihood function} \times \textit{Prior distribution}$$

Patz and Junker (1999) provided the MCMC approach for IRT models and demonstrated an application using a 2PL IRT model. The MCMC approach described in Patz and Junker, estimates both item parameters and person parameters jointly as in the traditional JMLE. MCMC estimation can be used for MixIRT models as well as IRT models (Baker, 1998;

Junker, 1999; Patz & Junker, 1999). Bolt et al. (2001, 2002), Wollack et al. (2003), and Cho et al. (2013) have estimated the MRM using WinBUGS. Additionally, Yamamoto and Everson (1997) estimated the HYBRID model using MCMC methods in WinBUGS. A Mix3PL model was also applied by Cohen and Bolt (2005) using code written in WinBUGS.

A more detailed description of MRM application with MCMC can be found in Cho et al. (2013). Bayesian estimation of MRM requires the definition of likelihood function and prior distributions for parameters. First step is to provide a likelihood function as shown in Equation 2.32. Assume that the responses of the n individuals to the test items are independent, the likelihood function of a MRM can be written as

$$L(g,\theta_j)=\prod_{i=1}^{I}\prod_{j=1}^{J}\left\{\left[\sum_{g=1}^{G}\pi_g P(y_{ijg}=1|g,\theta_{jg})\right]^{x_{ij}}\left[\sum_{g=1}^{G}\pi_g(1-P(y_{ijg}=1|g,\theta_{jg}))\right]^{1-x_{ij}}\right\}^{\zeta_{jg}^{t}}, \tag{2.32}$$

where x_{ij} is response vector coded with zero for a wrong answer and one for a correct answer and $\zeta_{jg}^{t}=1$ if examinee j is in class g at iteration t and 0 otherwise.

As mentioned earlier, prior distributions must be specified for all unknown parameters in order to estimate these parameters. In the context of the MRM, estimates of ability parameters, θ_{jg}, item difficulty, β_{ig}, group membership, g, and mixture probabilities, π_g, are of interest. For the Mix2PL model, α_i, an item discrimination parameter is also estimated. For the Mix3PL model a guessing parameter, γ_i, is also estimated. Selection of priors for these parameters plays an important role in Bayesian estimation. Priors can be either uninformative or informative and can also use information from previous results. A normal prior is often used for the distributions of ability and item difficulty. Cho et al. (2013) note that use of normal priors has been found to yield reasonable estimates of the item parameters in MixIRT models. Group membership has a multinomial distribution composing of one draw on g categories with mixture probabilities (e.g., $g\sim$Multinomial(1, π_g); Congdon, 2003; McLachlan & Peel, 2000). Dirichlet prior and a Dirichlet process with stickbreaking prior are described by McLachlan and Peel (2000) as conjugate priors for mixture probabilities.

Suitable conjugate priors produce effective Bayesian estimation by yielding proper posterior densities (McLachlan & Peel, 2000). Cho et al. (2013) found no substantial differences from use of either of these two priors in the estimation of the MRM. The MCMC algorithm is used to sample class membership parameter for each examinee at each stage of the Markov chain (Bolt et al., 2002; Cho et al., 2013). For instance, in a two-class MRM, class membership parameters, $g = 1, 2$ are sampled for each examinee j. Then group membership parameter of each examinee is determined as the mode of gs across iterations following burn-in (Cho et al., 2013).

2.5 Model Selection Issues

Determining the number of classes in the data is another important issue in MixIRT models. Several fit statistics are available for the selection of appropriate IRT model. In general, information criterion indices or likelihood ratio (LR) test statistics can be used within the IRT context. The former is appropriate for nonnested IRT models while the latter can be used but only for nested IRT models (Li et al., 2009). Only information criterion indices were used in this study because MixIRT models are not nested models (Li et al., 2009).

2.5.1 Information Criteria

Information criteria are based on some form of penalization of the likelihood function. Two of the more widely used information criteria are AIC and BIC. There are also several extensions of these two main indices such as CAIC and SABIC. These indices are described in Akaike (1974), Schwarz (1978), and Li et al. (2009). Typically, these information criterion indices are used to compare different model solutions of the same data when the ML estimates of the parameters are obtained. Smaller values of information criterion indices indicate better fit. However, they may provide different solutions to the same data due to differences in the penalty function applied to likelihood. AIC and BIC indices are discussed below. AIC can

be calculated as follows:

$$\text{AIC} = -2\log L + 2d, \tag{2.33}$$

where L is the likelihood function and d is the number of estimated parameters calculated as follows:

$$d = m * I * j + 2 * j - 1, \tag{2.34}$$

where m can have values from 1 to 3 for the MRM, Mix2PL and Mix3PL IRT models, respectively, I denotes the number of items, and j is the number of latent classes. For example, $j = 2$ is used for a two-class MixIRT solution.

As shown in the Equation 2.33, $2d$ is used as a penalty for over parameterization in the AIC index. One problem with AIC is that it does not take the sample size into account and tends to select more complex models (Li et al., 2009). The lack of penalization for sample size leads inconsistency in performance of AIC (Tofighi & Enders, 2007). However, the CAIC applies a penalty function that uses both the number of parameters and the sample size:

$$\text{CAIC} = -2\log L + d[\ln(N) + 1]. \tag{2.35}$$

In order to account for sample size, the BIC information criterion can be used. The BIC has been found to be somewhat more accurate than AIC for selection of MixIRT models (Li et al., 2009; Preinerstorfer & Formann, 2011). BIC tends to select simpler models than AIC due to the inclusion of the number of parameters in the penalty function. BIC can be calculated as follows:

$$\text{BIC} = -2\log L + d * \ln(N), \tag{2.36}$$

where L is the likelihood of the estimated model with d free parameters and $\ln(N)$ is the natural log of the total sample size N. As can be seen in Equation 2.36, the penalty function for BIC includes the number of estimated parameters and the natural log of the sample size. The sample size adjusted BIC uses $[N^* = (N + 2)/24]$ instead of N. This is proposed by Sclove (1987) to reduce the penalty on the sample size.

AIC and BIC as described here are based on a likelihood estimated using MLE. Use of Bayesian estimation requires a different likelihood function (i.e., posterior mean of the deviance). The MLE based deviance value ($-2L$) will be replaced in this study with the posterior mean of the deviance $\overline{D(\xi)}$ as obtained via MCMC estimation (Congdon, 2003) where ξ represents all estimated parameters in the model. Likelihood function was monitored at each iteration and final likelihood value was taken as the posterior mean of the deviance. As shown in Li et al. (2009), AIC and BIC formula can be re-parameterized as $\overline{D(\xi)}+2d$ and $\overline{D(\xi)}+d^*\log(N)$, respectively. Both of these indices will be reported in the current study.

2.6 Latent Non-normality

Distribution of the latent variable is typically assumed to be normal in most IRT applications particularly when unidimensional IRT models are fitted with MMLE. The assumption of normality of the latent variable, however, has been questioned in the literature (Andersen & Madsen, 1977; Bock & Aitkin, 1981; Mislevy, 1984; Thissen, 1991; Woods, 2004, 2006, 2007). Consequently, it may be unrealistic to assume normality for all latent variables of interest. If the latent variable is not normally distributed, assuming it is normal may have a biasing effect on parameter estimates (Kirisci, Hsu, & Yu, 2001; van den Oord, 2005; Zwinderman & van den Wollenberg, 1990).

It has been suggested that the distribution of latent variable should be checked before applying unidimensional IRT with MMLE (Woods, 2008). Several methods (described below) have been proposed for dealing with latent non-normality in IRT including alternative parametric form (Andersen & Madsen, 1977), the empirical histogram method (Bock & Aitkin, 1981; Mislevy, 1984), using Johnson curves (Thissen, 1991) and using Ramsay-curves method (RC-IRT; Woods, 2004).

In the empirical histogram approach, the latent ability distribution is approximated by a discrete distribution on a finite number of equally spaced points (Bock & Aitkin, 1981).

There is no restriction on the shape of the latent distribution in this approach. The computer program BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003) can be used for this approach. Johnson curves provide another approach. This approach is implemented in the computer software MULTILOG (Thissen, 1991). This method estimates the distributional parameters from the data by characterizing the latent trait distribution via the Johnson system of curves (van den Oord, 2005; Woods, 2007). The Ramsay curve IRT method was designed to detect violations of normality and to obtain corrected estimates of item parameters when the distribution of latent variable is non-normal (Woods, 2004). The latent ability distribution and item parameters are estimated simultaneously with this approach (Woods, 2007). Simulation studies showed that estimation with the Ramsay curve method resulted in more accurate estimates than normality-based estimation when the ability distribution is skewed (Woods, 2004, 2006).

Latent non-normality problem has not been well studied in MixIRT models. Xu and Jia (2011) estimated mixed Rasch and mixed 2PL models with generalized skewed normal and standard normal distributions. Based on the results from these two distributions, Xu and Jia demonstrated that the item parameter estimates and the ability distributional statistics were similar in all conditions between two distributions. However, the standard errors of item parameters were found to be larger in the skewed distribution condition.

Chapter 3

METHOD

As noted above, the purpose of this study is to investigate the effect of non-normal ability distributions on the number of latent class extracted in MixIRT models. To this end, this study investigated the following two research questions:

1. Is the accuracy of detection of latent classes affected by using a normal prior on ability parameters when the latent ability distribution is non-normal?

2. What effect do skewness and kurtosis have on the extraction of latent classes for MixIRT models?

Analyses to investigate these two research questions were described in more detail in the sequel. In the next sections of this chapter, two simulation studies and an empirical study are described.

3.1 Fixed Conditions

3.1.1 Test Length

A test length of 28 items was used since this is similar to the number of items on the state test used for obtaining estimates of item parameters for the simulation study. Tests of this length can be considered to be an average length. In the second condition, the test length of 10 items was used in order to investigate the sensitivity of a short test to latent non-normality.

3.1.2 Sample Size

Li et al. (2009) report results suggesting samples of 600 may be appropriate 15- or 30-item tests for the MRM. Li et al. suggest for a 15-item test that samples of 600 would be sufficient for a 1- to 4-group model for both the Mix2PL and Mix3PL IRT models. Cho et al. (2013) suggest a sample size greater than 360 can be used when estimating a MRM, and Cho, Cohen, Kim, and Bottge (2010) report accurate model selection with samples as small as 100 for a 2-class MRM. Cohen and Bolt (2005) successfully applied Mix3PL IRT model with a sample size of 1,000. A sample of this same size also was used in Bolt et al. (2002). It is evident that the more complex models that are estimated, the larger the sample size will need to be to estimate those models. A sample size of 600 was proposed as the small sample condition for the present study.

In the second sample size condition, 2,000 examinees were simulated. This represents a large sample size condition in this study. Rost (1990) analyzed an MRM with a sample size of 1,800 and Samuelsen (2005) used a sample size of 2,000 examinees. A summary of the four factors included as simulation conditions is presented in Table 3.1.

Table 3.1: Summary of the Manipulated Conditions in Simulation Study 1

Ability distribution	Test length	Sample size	Replications
	10	600	50
Bimodal Symmetric		2,000	50
	28	600	50
		2,000	50
	10	600	50
Normal		2,000	50
	28	600	50
		2,000	50
	10	600	50
Platykurtic		2,000	50
	28	600	50
		2,000	50
	10	600	50
Skewed		2,000	50
	28	600	50
		2,000	50
	10	600	50
Uniform		2,000	50
	28	600	50
		2,000	50

3.1.3 Generating Item Parameters

Generating item parameters for two simulation study were provided the item parameters obtained from Rasch, 2PL, and 3PL model estimates of Florida Comprehensive Assessment Test (FCAT; Florida Department of Education, 2002) mathematics test for Grade 9. Item parameters for FCAT data were estimated using MULTILOG 7.03 (Thissen, 2003). In using MULTILOG, all default options were used and number of options was set to four. Only 28 items in the FCAT Grade 9 data set were multiple choice items. These were coded dichotomously (0 = incorrect; 1 = correct). The remaining 22 items were either short answer or gridded response items. Only the item parameter estimates from the complete set of dichotomous items were used as generating parameters in the 28-item test length condition. The first ten dichotomous items from FCAT test data were used for the 10-item test. Estimated item parameters for these three models are presented in Tables 3.2 and 3.3 (a - slope parameter, b - threshold parameter, and c - guessing parameter).

Table 3.2: Item Parameters Used for Data Generation for 10-Item Condition

	RM	2PL Model		3PL Model		
Item	b	a	b	a	b	c
1	−1.83	0.91	−1.84	0.91	−0.37	.23
2	−0.07	0.93	−0.07	1.17	0.69	.30
3	−0.15	1.21	−0.13	1.23	0.39	.24
4	0.90	0.84	0.94	0.91	1.23	.16
5	−0.38	0.94	−0.37	0.66	−0.06	.12
6	−0.59	1.14	−0.51	0.75	−0.37	.06
7	0.98	0.76	1.14	0.76	1.38	.14
8	0.51	1.06	0.45	1.28	0.88	.22
9	0.99	0.34	2.37	0.87	1.67	.28
10	0.19	1.27	0.15	1.05	0.46	.14

Table 3.3: Item Parameters Used for Data Generation for 28-Item Condition

	RM	2PL Model		3PL Model		
Item	b	a	b	a	b	c
1	−1.72	1.05	−1.66	1.45	−0.45	.20
2	−0.09	0.88	−0.10	1.66	0.76	.31
3	−0.16	1.24	−0.16	1.60	0.40	.24
4	0.81	0.72	1.04	0.62	1.35	.19
5	−0.37	0.93	−0.39	0.74	0.05	.16
6	−0.57	1.28	−0.50	1.35	−0.34	.06
7	0.91	0.72	1.16	1.31	1.40	.15
8	0.45	1.07	0.42	1.32	0.88	.22
9	0.91	0.38	2.08	1.47	1.67	.26
10	0.16	1.27	0.12	1.25	0.48	.15
11	0.69	0.67	0.95	1.42	1.34	.25
12	0.42	0.94	0.43	1.26	0.93	.23
13	0.93	0.69	1.26	1.61	1.35	.22
14	1.22	0.98	1.24	1.17	1.29	.14
15	0.31	0.94	0.32	1.66	0.81	.20
16	1.19	0.92	1.25	1.38	1.30	.16
17	0.27	1.18	0.23	1.47	0.72	.22
18	−1.54	1.61	−1.15	1.59	−1.15	.03
19	−0.39	1.69	−0.32	1.43	−0.15	.06
20	−0.41	1.46	−0.35	0.77	−0.03	.14
21	−0.34	1.01	−0.34	1.27	0.12	.17
22	−0.30	1.22	−0.28	1.64	0.46	.32
23	0.18	1.87	0.08	1.45	0.30	.09
24	0.09	0.76	0.13	1.33	0.97	.32
25	0.10	0.70	0.15	1.01	0.72	.18
26	−0.31	1.01	−0.31	1.12	−0.09	.08
27	−0.33	0.91	−0.35	0.93	−0.32	.00
28	−0.47	1.43	−0.39	1.49	−0.01	.17

3.1.4 Number of Replications

Each condition was simulated 50 times. This amount of replication is large for MCMC studies with very complex models, such as those used in this study. This also may be viewed as small when compared to estimation for less complex models. The number of replications was chosen to provide data for analyses that should be capable of detecting the effects of latent nonnormality, if any exist. There is as yet no data on the extent of these effects or the effect sizes to expect for the recovery analyses.

In addition, MCMC estimation for the complex models used in this study takes much longer than for estimation of much simpler unidimensional, single-class IRT models. The amount of computational time that is required for MCMC estimation has been found to vary directly with model complexity and with data characteristics such as the number of items and sample size (e.g., Li et al., 2009). In this study, the MCMC run for MRM analyses required times from 1 hour to 5 hours based on the sample size and test length conditions. The amount of time that was required for the Mix2PL model analyses ranged from 2 to 13 hours. The Mix3PL model analyses took from 6 to 36 hours to complete for the different sample sizes and test lengths conditions simulated. All of these analyses for Simulation Study 1 were completed using a Windows operating system on a 2.00 GHz server blade with a Quad-Core Intel Xeon E5405 processor with 5GB RAM. Analyses for Simulation Study 2 were done using a Linux cluster comprised of computer nodes with 4, 6, 8, and 12-core processors.

3.2 Simulation Study 1

To investigate the effect of non-normality, a Monte Carlo simulation study was conducted to model the data, using the item parameters from FCAT test. The following conditions were simulated: Sample size (600 and 2,000 examinees), test length (10 and 28 items), and five ability distributions (bimodal symmetric, normal, platykurtic, skewed, and uniform). Data

were simulated for each of the three dichotomous IRT models × 3 MixIRT models for fitting × 2 classes for fitting (one- and two-classes) × 2 sample sizes × 2 test lengths × 5 ability distributions = 360 conditions. Fifty replications were simulated for each condition.

3.2.1 Distribution Conditions

Five distributions were considered for the ability parameter in the first simulation study: normal, skewed, platykurtic, bimodal symmetric and uniform. In the normal condition, the ability distribution was randomly drawn from a standard normal distribution with unit variance (i.e., $N(0, 1)$). This condition was added for completeness as it is used by much of the software available for estimating IRT models. Skewed and platykurtic data were generated using the power method proposed by Fleishman (1978). The power method is a procedure used to generate non-normality for Monte Carlo studies based on the priori known parameters such as skewness and kurtosis. There are several other methods to generate non-normal distributions including the addition of outliers to the data, the use of some of the known distributions (e.g., chi-square, exponential, and Cauchy distributions), and transformation to unknown non-normality using a skewing function (Fleishman, 1978). Although these techniques can be used to generate a non-normal data, it is not possible to obtain the skewness and kurtosis values that we wish in advance. On the other hand, Fleishman's power method provides the non-normal distributed data with the desired skewness and kurtosis values. One limitation with this method is that it cannot provide all the values within parameter space. The obtainable values can be identified using the following formula:

$$skew^2 = 0.0629 * kurtosis + 0.0717. \tag{3.1}$$

First, normal theta values were generated and then skewed and platykurtic data were generated from these data using constants given in Fleishman (1978). The power method

was used to generate skewed and kurtotic data by changing four constants (a, b, c, and d) in the polynomial given below:

$$Y = a + bX + cX^2 + dX^3, \tag{3.2}$$

where X is a random variate distributed normally with zero mean and unit variance, and a, b, c, and d are weights used to change distribution of Y. Actually, only three constants (b, c, and d) are required for transformations as the constant a can be shown to be equal to $-c$ (Fleishman, 1978).

Skewness and kurtosis values were 0.75 and 0.0 for skewed data and 0.0 and -0.75 for platykurtic data condition in this simulation study. These values were selected to represent the typical non-normality situation. Typical non-normality refers to distributions with skewness less than 0.8 and kurtosis between -0.6 and 0.6 (Pearson & Please, 1975). The constants used to obtain these two conditions are provided in the power weights table in Fleishman (1978). To obtain skew = 0.75 and kurtosis = 0.0, $a = -0.1736$, $b = 1.1125$, $c = 0.1736$, and $d = -0.0503$; to obtain skew = 0.0 and kurtosis = -0.75, $a = 0.0$, $b = 1.1336$, $c = 0.0$, and $d = -0.0467$.

In the uniform condition, ability parameters were randomly drawn from Uniform(-2, 2). The ability parameters for bimodal symmetric condition were randomly drawn from combination of two normal distributions $N(-1.5, 1)$ and $N(1.5, 1)$. All of the conditions were generated using R program. Graphical representations of the four non-normal conditions are presented in Figure 3.1. A normal distribution curve is superimposed on each figure for reference.

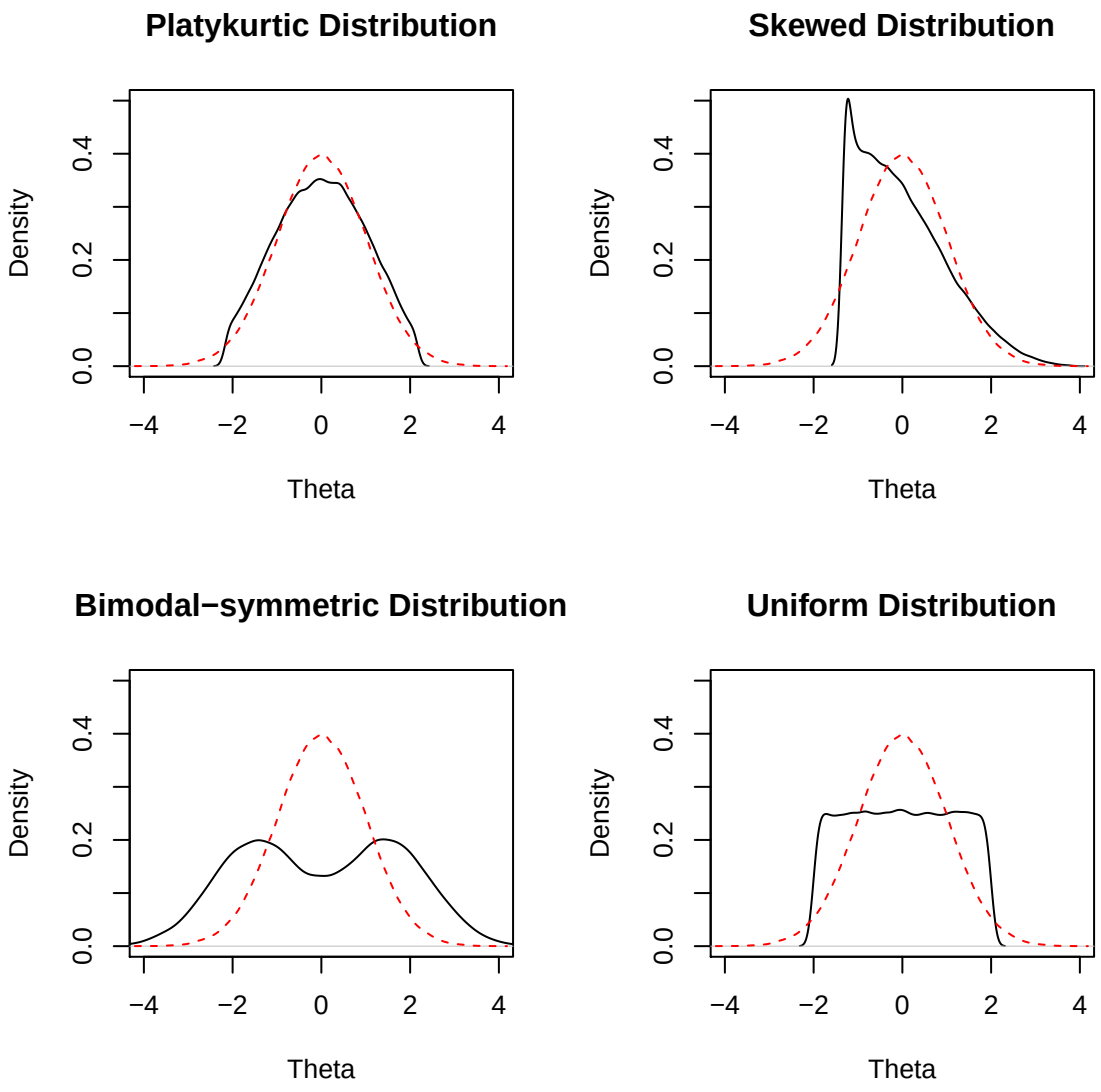

Figure 3.1: Ability distributions for simulation study 1.

3.3 Simulation Study 2

Simulation Study 2 focused on non-normal data with various combinations of skewness and kurtosis values. As mentioned above, for some testing conditions, non-normal ability distributions are more likely to occur than are uniform and bimodal ability distributions in

educational and psychological data (Gregoire & Driver, 1987; Micceri, 1989). The purpose of this simulation study was to investigate the values of skewness and kurtosis that may cause over-extraction of latent classes. Simply, we tried to find a safe region for correct identification of number of latent class in skewed and kurtotic data. The power method proposed by Fleishman (1978) that was used in Study 1 also used in Study 2.

Table 3.4: Power Method Weights Adapted from Fleishman (1978)

Condition	Skewness	Kurtosis	b	c	d
Condition 1	1.50	4.00	0.9296	0.3994	−0.0364
Condition 2	1.50	3.50	0.8869	0.2327	0.0187
Condition 3	1.50	2.50	0.9920	0.3452	−0.0418
Condition 4	1.00	3.50	0.8029	0.1221	0.0573
Condition 5	0.50	3.50	0.7689	0.0564	0.0708
Condition 6	0.00	3.50	0.7590	0.0000	0.0747

3.3.1 Distribution Conditions

Various combinations of skewness and kurtosis values were simulated to investigate problematic distributional parameters in terms of over extraction. Six different combinations of skewness and kurtosis presented in Table 3.4 were simulated. Graphical representations of the six distributions are presented in Figure 3.2. A standard normal distribution curve was superimposed for reference on the graphs. Two sample sizes (600 and 2,000 examinees) were simulated for two test lengths (10 and 28 items). The same generating values used for item parameters in Study 1 were used in Study 2. In summary, the following factors were simulated for Study 2: 3 IRT models for data generation × 3 MixIRT models for fitting × 2 classes for fitting (one- and two-classes) × 2 sample sizes × 2 test lengths × 6 ability distributions = 432 conditions. A total of 50 replications were simulated for Study 2.

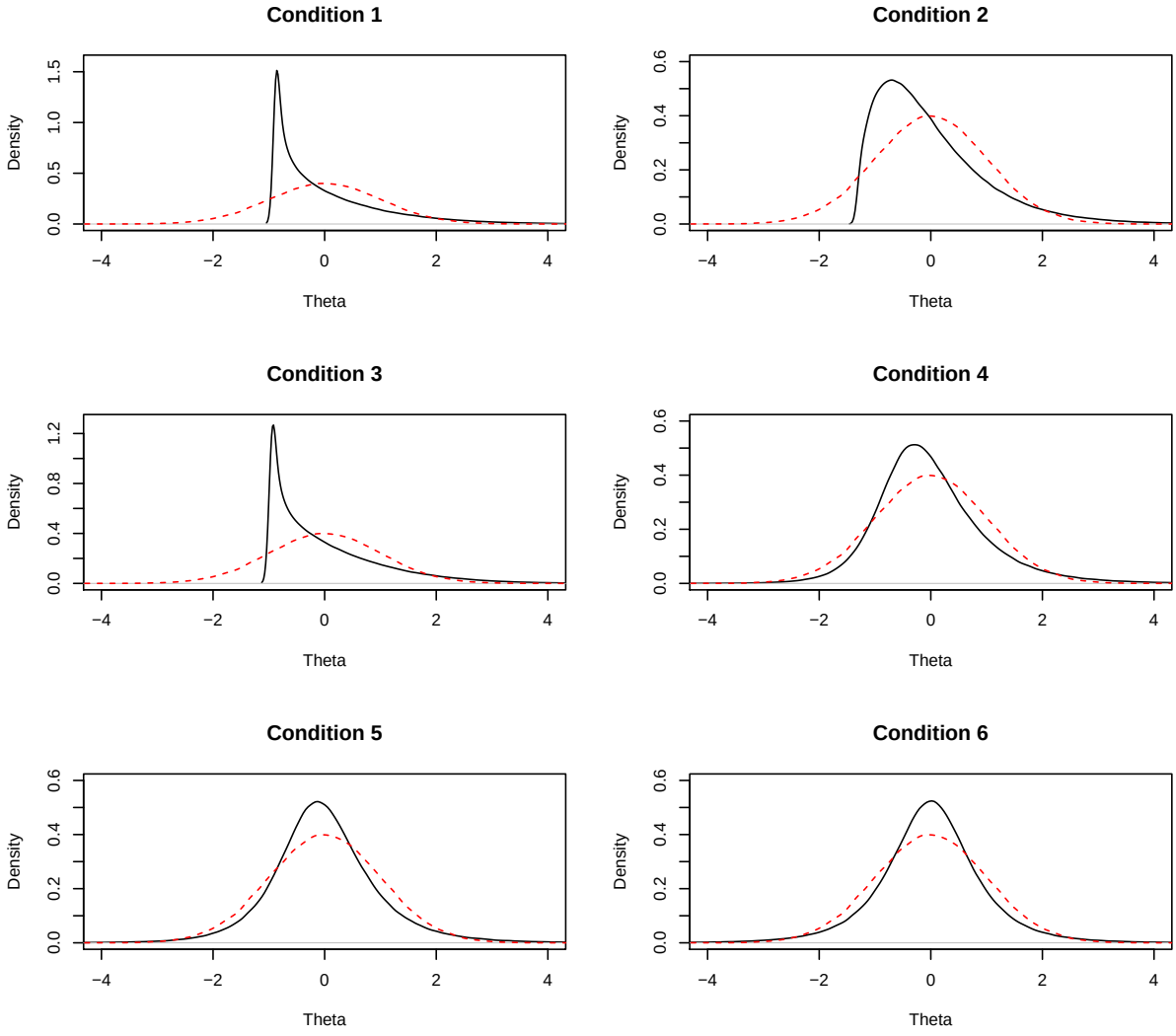

Figure 3.2: Ability Distributions for Simulation Study 2

3.4 Unidimensionality Analyses

It is important to determine if unidimensionality holds before any unidimensional IRT modeling is applied. Unidimensionality of the generated data sets were examined before estimation of MixIRT models. It is often possible to have a single dominant dimension (i.e., essential dimensionality) rather than a single dimension (i.e., strict dimensionality) in IRT

application. Reckase (1979) suggests that if the first component explains at least 20% of the variance, the test can be considered to be unidimensional.

There are several methods proposed to assess test dimensionality for dichotomous type items. Most of these methods are based on item factor analysis (IFA) models within the context of structural equation modeling (SEM) and IRT. The full-information maximum-likelihood (FIML) estimation (Bock, Gibbons, & Muraki, 1988), the algorithm in the software package LISCOMP (Muthén, 1978), nonlinear factor analysis (McDonald, 1982), and factor analysis of the tetrachoric correlations between all item pairs (Knol & Berger, 1991) are the most popular IFA-based techniques. These approaches use different estimation algorithms and data types. For instance, the FIML estimation method analyzes the entire item response pattern while the analyses in other three approaches are based on bivariate information. Nonparametric approaches for dimensionality assessment are also available such as the algorithm in the computer software DIMTEST (Nandakumar & Stout, 1993) and in the software DETECT (Zhang & Stout, 1999).

There are a number of software packages designed for both parametric and nonparametric approaches including DIMTEST, Mplus, NOHARM (Fraser & McDonald, 1988), and TESTFACT (Wilson, Wood, & Gibbons, 2003). While DIMTEST software is designed for nonparametric approach other three programs can be used for IFA-based approach. TESTFACT can handle guessing parameter, but Mplus does not have any option for handling the guessing parameter in the 3PL IRT model.

Unidimensionality of each simulated data set was examined before implementing MixIRT model analyses. Since TESTFACT can be used to model guessing, it was used with FIML estimation to perform an exploratory factor analysis (EFA) on each dataset in this study. Dimensionality assessment can be done in an exploratory way using a chi-square statistic provided by TESTFACT. This approach is based on comparison of chi-square statistics between two analyses with different components (e.g., single factor vs. two-factor solution).

The proportion of variance accounted for is another approach that can be used to investigate dimensionality. The second approach is more flexible than chi-square approach and can be used for determining the dominant factor based on use of Reckase's 20% criteria. These two approaches with TESTFACT software were used in this study.

3.5 Model Specifications for MCMC Estimation

In this study, the RM, 2PL, and 3PL models were used to generate the simulation data. Each model was generated to have one class. The data generated by the RM were fitted with the MRM and the data generated by 2PL and 3PL models were fitted with Mix2PL and Mix3PL models, respectively. Given that our purpose is to see whether a two-class solution (i.e., a spurious class situation) would fit where a one-class model was simulated, each MixIRT model was fitted with one- and two-class solutions.

WinBUGS software was used for estimation of models for each of the generated data sets. Priors for parameters can be either fixed values from a known data set (Cohen & Bolt, 2005) or some known distributions (Bolt et al., 2002; Cho et al., 2013; Li et al., 2009). Typically, a standard normal distribution $N(0, 1)$ was used for item difficulty and discrimination parameters in MCMC estimation of MixIRT models (Bolt et al., 2002; Cho et al., 2013; Cohen & Bolt, 2005; Li et al., 2009). The beta distribution is generally used as a prior for the guessing parameter in a mixture 3PL model (Li et al., 2009) as well as in a 3PL model (Johnson & Albert, 1999). These mildly informative priors were used to obtain more stable estimates (Cho et al., 2013). In addition to these priors, a non-informative prior with an equal likelihood for each class was used for mixing proportions in two-group analyses in MixIRT estimations.

As mentioned above, the ARS algorithm was used for MCMC analyses. When the target densities are log-concave, as expected in our analyses, the ARS method is automatically implemented in the WinBUGS software (Patz & Junker, 1999). Typically, there are three

parts required in WinBUGS code to do Bayesian estimation. These include model specification, specifications of priors, and specification of initial values and data. After specifying the model in terms of the likelihood and prior distributions, initial values can be provided either by the user or generated randomly by WinBUGS. For the mixing proportions, .5 was used as weights. This treats both groups equally. The starting values for all other parameters were randomly generated using the WinBUGS software. Likelihood value was specified to compute AIC and BIC values.

The following priors and hyper-priors were used for the MRM:

$$\begin{aligned}
\beta_{ig} &\sim \text{Normal}(0,\ 1) \\
\theta_j &\sim \text{Normal}(\mu(\theta), 1) \\
\mu(\theta)_g &\sim \text{Normal}(0,\ 1) \\
g_j &\sim \text{Bernoulli}(\pi_1, \pi_2) \\
(\pi_1, \pi_2) &\sim \text{Dirichlet}(.5,\ .5),
\end{aligned}$$

where θ_j represents the ability parameter for examinee j, β_{ig} is the difficulty parameter of item i within class g, and π_1 and π_2 are class membership parameters. Estimates of the mean and standard deviation for each latent class, μ_g and σ_g, can also be estimated via MCMC. As recommended by Bolt et al. (2002), σ_g was fixed at 1 for both groups. A Dirichlet distribution was used as the prior for π_g for the two-group models. Two additional priors were used in Mix2PL and Mix3PL analyses:

$$\begin{aligned}
\alpha_{ig} &\sim \text{Normal}(0,\ 1)I(0,) \\
\gamma_{ig} &\sim \text{Beta}(5,\ 17).
\end{aligned}$$

Subscript g can be dropped for the one-group analysis. Detailed WinBUGS code is presented in the appendix.

3.5.1 Convergence Diagnostics

Determining model convergence is another important issue that needs to be resolved with MCMC. Lack of convergence may lead to imprecise parameter estimates and, therefore, to false inferences. Since the starting values may have an effect on the parameter estimates, a number of initial iterations are typically discarded in MCMC estimations. These are referred to as burn-in iterations. Final posterior estimates are often obtained using the remaining post-burn-in iterations. Results from previous MixIRT studies showed that a large number of burn-in and post-burn-in iterations were needed to obtain stable posterior distributions for all estimated model parameters. Li et al. (2009) found that 2,500 burn-in iterations were enough for the MRM and Mix2PL model analyses and 6,000 iterations for the Mix3PL model analyses. Cho et al., (2013) found 7,000 iterations to be needed for burn-in and 8,000 iterations for post-burn-in for the MRM. The number of such iterations depends on a number of factors, including test length, sample size and number of parameters to be estimated. Several methods have been proposed for convergence diagnostic in MCMC estimations (Cowles & Carlin, 1996). The convergence diagnostics by Gelman and Rubin (1992) and Raftery and Lewis (1992) are currently the most popular methods (Cowles & Carlin, 1996). Two R packages, Bayesian Output Analysis (BOA; Smith, 2007) and convergence diagnosis and output analysis for MCMC (CODA; Plummer, Best, Cowles, & Vines, 2006), can be used to conduct convergence assessment. In addition to these two methods, trace plots, density plots and Monte Carlo error statistic can be considered.

For complex models with many parameters, it is impractical to check convergence for every parameter, so the monitoring of a most relevant parameter would be more practical. Item difficulty parameter estimates were monitored in this study in order to help with determining the number of burn-in and post burn-in iterations. Only data sets with the large sample size (i.e., 2,000 simulated examinees) and the long test (i.e., 28 items) were used for convergence assessment. The number of burn-in and post burn-in iterations were determined

by examining the Gelman-Rubin convergence statistics. As an example of this convergence method, Figure 3.3 presents Gelman-Rubin shrink factor plots for Item 5 of the Rasch, 2PL, and 3PL model data sets.

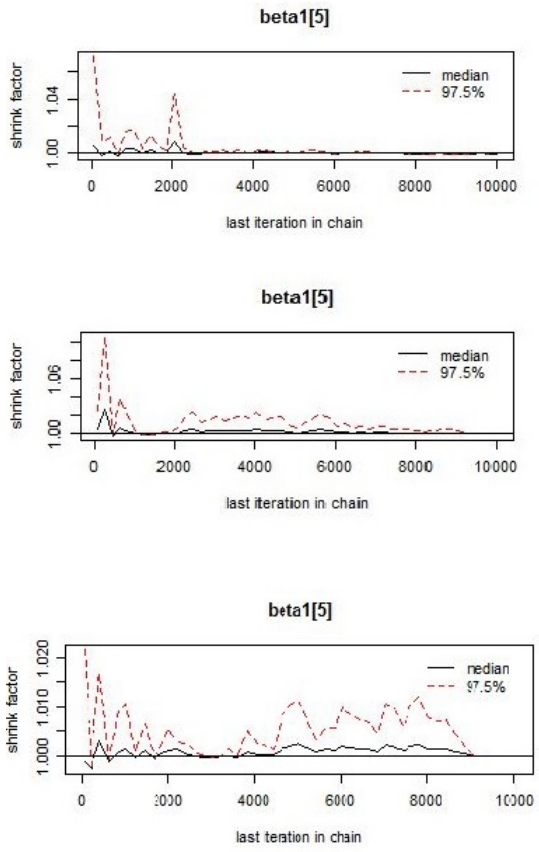

Figure 3.3: Gelman-Rubin Shrinkage Factor Plots for MRM, Mix2PL, and Mix3PL models, respectively from top to bottom.

Calculation of Gelman-Rubin convergence statistics requires at least two chains. So, MCMC analyses for each condition in this study were done with two parallel chains. As shown in the upper plot of Figure 3.3, two chains appear to have reached a stationary distribution, converging around 1, after about 3,000 iterations for the MRM. This can be seen by examining the within and between chain variance for difficulty parameter estimate.

Similarly, the middle plot in Figure 3.3 shows that two chains have reached a stationary distribution for the Mix2PL model, converging to around 1 at about 8,000 iterations. In the lower plot of Figure 3.3, the Gelman-Rubin statistic for the Mix3PL converges around 1 after about 9,000 iterations.

Along with plots for individuals items, the examples of Gelman-Rubin shrink factor plots are given in Appendix A for (1) an MRM data set with 28 items and 2,000 examinees, (2) a Mix2PL data set with 28 items and 2,000 examinees and (3) a Mix3PL data set with 28 items and 2,000 examinees. Convergence for most items required only a small number of burn-in iterations, but a few items required more iterations. Based on preliminary results, a conservative number of burn-in iterations was selected for three different model analyses. Six thousand iterations were discarded as burn-in and 6,000 were used as post-burn-in iterations for the MRM conditions. For the Mix2PL model conditions, 7,000 burn-in iterations and 7,000 post burn-in iterations were used, and 9,000 burn-in iterations and 9,000 post burn-in iterations were used in all Mix3PL IRT model conditions. The results of post burn-in iterations are discussed in the Results chapter along with details.

3.6 Evaluation Criteria

Recovery of item parameters was assessed using root mean square errors (RMSEs) and bias between the generating parameter and the parameter estimate. Lower RMSE values indicate better estimation accuracy. Similarly, the smaller the bias, the more accurate the parameter estimates are in the considered condition. A zero value of bias indicates the estimation results in unbiased parameter estimates. The computational formulas for RMSE and bias for difficulty parameters are presented below:

$$\text{RMSE}_{(\beta_i)} = \sqrt{\frac{\sum_{i=1}^{I}\sum_{r=1}^{R}(\beta_i - \hat{\beta}_{ir})^2}{RI}} \tag{3.3}$$

$$\text{Bias}_{(\beta_i)} = \frac{\sum_{i=1}^{I}\sum_{r=1}^{R}(\beta_i - \hat{\beta}_{ir})}{RI}, \tag{3.4}$$

where β_i and $\hat{\beta}_{ir}$ are generating and estimated item difficulty parameters for item i, respectively, I $(i = 1, ..., I)$ denotes the number of items and R $(r = 1, ..., R)$ is the number of replications. In addition to examination of difficulty parameters, it is necessary to evaluate the recovery of discrimination and guessing parameters in 3PL model. Same formula can be used for these parameters as shown below. RMSE and bias values for estimates of discrimination parameters:

$$\text{RMSE}_{(\alpha_i)} = \sqrt{\frac{\sum_{i=1}^{I}\sum_{r=1}^{R}(\alpha_i - \hat{\alpha}_{ir})^2}{RI}} \tag{3.5}$$

$$\text{Bias}_{(\alpha_i)} = \frac{\sum_{i=1}^{I}\sum_{r=1}^{R}(\alpha_i - \hat{\alpha}_{ir})}{RI}, \tag{3.6}$$

where α_i denotes generating item discrimination parameters for item i and $\hat{\alpha}_{ir}$ represents estimated item discrimination parameters for item i and replication r. RMSE and bias values for guessing parameters can be estimated in the same way.

$$\text{RMSE}_{(\gamma_i)} = \sqrt{\frac{\sum_{i=1}^{I}\sum_{r=1}^{R}(\gamma_i - \hat{\gamma}_{ir})^2}{RI}} \tag{3.7}$$

$$\text{Bias}_{(\gamma_i)} = \frac{\sum_{i=1}^{I}\sum_{r=1}^{R}(\gamma_i - \hat{\gamma}_{ir})}{RI}, \tag{3.8}$$

where γ_i and $\hat{\gamma}_{ir}$ are generating and estimated item guessing parameters for item i, respectively. The estimates and the generating parameters should be placed on the same scale before calculating RMSE and bias (Kolen & Brennan, 2004). Thus, in this study, the estimated parameters were placed on the scale of the generating parameters using the following transformation:

$$\hat{\beta}_i^* = \hat{\beta}_i - (\bar{\hat{\beta}}_T - \bar{\beta}_B), \tag{3.9}$$

where B represents base scale (generating parameters) and T represents target scale (estimated parameters). $\hat{\beta}_i^*$ denotes the equated difficulty for the estimated item i, $\hat{\beta}_i$ is the estimated difficulty parameter for item i, and $\bar{\hat{\beta}}_T$ and $\bar{\beta}_B$ denote the means of the item difficulty estimates for the simulated data set and generating parameters, respectively.

As mentioned earlier, AIC and BIC values were calculated for each run. Since one-class data sets were generated, these fit indices were used to determine whether a multiple-class (2-Class) model had better fit than a single class model. A percentage of correct detection of simulated latent classes was calculated based on AIC and BIC for each condition. The percentage of correct identification is defined as the proportion of correct positives in single-class detection.

3.7 Empirical Study

An empirical data set was analyzed to illustrate the impact of different ability distributions on the number of latent classes detected in a MixIRT application. The purpose of this real data analysis was to show how to interpret results for large-scale assessments in the presence of latent non-normality in the data. MixIRT models were applied to an 8th grade mathematics data set (with non-normal distribution of ability) collected by the Trends in International Mathematics and Science Study (TIMSS; Foy, Arora, & Stanco, 2013) in 2011. TIMSS administers reliable and timely tests on the mathematics and science achievement to 4th- and 8th-grade students from more than 63 countries. TIMSS was first developed in 1995 by the International Association for the Evaluation of Educational Achievement. It has been administered every four years since then. Countries participating in the TIMSS program have the chance to compare their students achievement in mathematics and science to students of other participating countries. In addition to mathematics and science questions, TIMSS provides teacher and student questionnaires during the exam. The TIMSS test is scaled to have a mean of 500 and a standard deviation of 100. The latest administration of TIMSS was in 2011.

One of the highest performing countries, South Korea (henceforth, Korea), was selected for the analyses in the empirical study. Korea TIMSS 2011 8th grade mathematics test data was negatively skewed as most of the students taking the test score highly. The RCLOG

software (Woods, 2006b) was used to detect the skewness and kurtosis values of the latent distribution in this data set. As have been done in the two simulation studies, this skewed data set was analyzed with each of the three MixIRT models. The effect of the skewness on the parameter estimates and number of latent class were examined. The TIMSS assessment includes both multiple-choice and constructed-response items on the mathematics tests. In this study, only the multiple-choice items were used. The empirical data set was analyzed with one- to five-class solution MixIRT models to see whether the latent non-normality of the data set had an effect on the number of latent classes extracted. The description of the data set, descriptive statistics and the results from MixIRT model applications are described in the Results Chapter.

Chapter 4

RESULTS

The results of this study are presented in four sections in this chapter. The first section presents the preliminary results, including convergence assessment and unidimensionality analyses. The results of the simulation studies are presented in the next two sections which is followed by the results of empirical study in the last section.

4.1 Preliminary Analyses

4.1.1 Unidimensionality Assessment

The data sets generated for this study were designed to be unidimensional. To ensure that unidimensionality was achieved, additional analyses needed to be performed. Before proceeding to the MixIRT analyses, therefore, the dimensionality of each data set was first examined with the TESTFACT computer program (Bock et al., 1988). As mentioned earlier, the chi-quare difference test and the proportion of variance accounted for were used with TESTFACT. Dimensionality assessment with the chi-square statistic provided by TESTFACT can be performed through an exploratory factor analysis (EFA). The TESTFACT syntax was created to perform two EFAs on each data set. The number of factors extracted was set to one and two in order to implement the chi-square procedure. Additionally, guessing parameter values were set to .25 in TESTFACT for 3PL data sets. For each solution a chi-square fit statistic and eigenvalues were obtained. Based on these two chi-square fit statistics, the chi-quare difference test can be used to determine whether or not unidimensionality holds

in the data. As explained in the TESTFACT manual (du Toit, 2003), the chi-square statistic in TESTFACT can be used to test the difference between following two hypotheses. That is the test of following null hypothesis against the alternative hypothesis that is presented next.

H_0 : N-factor model provides an adequate fit to the data.

H_a : (N+1)-factor model provides an adequate fit to the data.

This chi-square statistic is simply the difference between the chi-square value under H_0 (i.e., the one-factor solution) and the chi-square value under H_a (i.e., the two-factor solution). Similarly, the degrees of freedom is obtained as the difference between the degrees of freedoms of the two chi-square statistics. Since a chi-square test statistic was used, the resulting difference value also has a chi-square distribution. Based on the hypotheses (above), a significant p-value would indicates that a two-factor solution provides a better fit to data than a one-factor solution.

The calculation of a chi-square test in TESTFACT, however, requires non-empty cells in the response vector with the number of unique responses equal to 2^k, where k is the number of items (Mislevy, 1986). Very small expected frequencies may have a negative effect on the reliability of chi-square test statistics (Mislevy, 1986). A three-step process was followed to obtain unidimensional data sets in this study. First, each data condition was screened for non-significant chi-square test values. Data sets that had a nonsignificant chi-square test value were retained as they were considered unidimensional. Second, data sets that had a significant chi-square statistic were re-examined in terms of proportion of variance accounted for values. The data sets with dominant factors (i.e., more than 20% proportion of variance) were retained for the analyses based on Reckase's (1979) 20% criteria. As mentioned above, the eigenvalues were computed from the tetrachoric correlation matrix in TESTFACT. Using the eigenvalues from TESTFACT output, the proportion of variance accounted for the first factor was computed for each data set. The data set

was considered undimensional in the case of eigenvalue analysis verified the dominant first factor. Third, data sets that did not meet either of these two criteria were dropped, and new data sets were generated until fifty unidimensional data sets were obtained for each condition.

```
TESTFACT OUTPUT FOR A UNIDIMENSIONAL DATA SET BASED ON PROPORTION OF VARIANCE

PHASE 6:  FACTOR ANALYSIS
NON-ADAPTIVE FULL-INFORMATION ITEM FACTOR ANALYSIS
 -------------------------------------------------------------
 DISPLAY   2.   THE POSITIVE LATENT ROOTS OF THE CORRELATION MATRIX

              1         2         3         4         5         6

     1     3.047328  1.071122  0.950484  0.862282  0.844695  0.745946

              7         8         9        10

     1     0.707823  0.673387  0.582313  0.514620
```

Figure 4.1: A Sample TESTFACT Output for Proportion of Variance Accounted for Approach

For most of the ten-item conditions, the chi-square statistic could not be calculated. Thus, only proportion of variance accounted for approach was available for evaluation of these conditions. For example, as shown in the sample analysis in Figure 4.1, the first factor explained 30% of all variation in the data. Thus, a dominant factor was assumed for this case and the data set was used for the MixIRT analyses.

4.1.2 Convergence Assessment

A sample assessment is presented for Item 1 (item difficulty = -1.72) to provide a brief example of what has been done in the convergence process. Figure 4.2 presents a Gelman-Rubin shrink factor plot, a trace plot, a density plot, and an autocorrelation plot for Item 1 for a data set generated with a Rasch model. These plots are commonly used to assess the convergence of an MCMC estimation. Gelman-Rubin shrinkage plot is used to show that multiple chains converges to same target of stationary distribution. Traceplot refers to a plot that shows the iteration number and the value of the parameter at each iteration. This plot allows us to understand if a chain reached to its stationary distribution and the chain is mixing well. That is , it helps us to determine if the number of burn-in is enough. The density plot is simply a histogram of the distributions of every draw. Converged chains produce a uni-modal (bell-shaped) density plot while non-convergence is observed in the presence of multimodal density plots. Autocorrelation plot shows the correlation between every draw. It is expected to observe a decay in this plot if the correlation between each draw decreases.

As described earlier, MCMC analyses for each condition were applied with two parallel chains to determine the number of burn-in iterations using Gelman-Rubin convergence statistics. As shown in the upper plot of Figure 4.2, the two chains appear to have reached a stationary distribution, converging close to 1, after 5,000 iterations based on within and between chain variance for the difficulty parameter of Item 1. So, a burn-in of 6,000 in iterations was adopted for this item. Another way of assessing convergence can be done by investigating the mixing between chains. That is, we can observe the strong effect of starting values in the case of poor mixing. Mixing of parallel chains indicates good convergence due to less influence of starting values. The history of the sampling is shown in the second graph in Figure 4.2. The chains begin at different points and mix with well each other. In addition, the plot of the kernel density appears to be normally distributed, and the autocorrelation plot shows a clear decay over the subsequent 25 iterations. This is the kind of decay that is

characteristic of good convergence. Finally, the difficulty parameter was estimated for Item 1 to be -1.68 with a Monte Carlo error of 0.0006438 which is less than 5% of the sample standard deviation (Spiegelhalter et al., 2003). From this, we conclude that the MCMC chain converged.

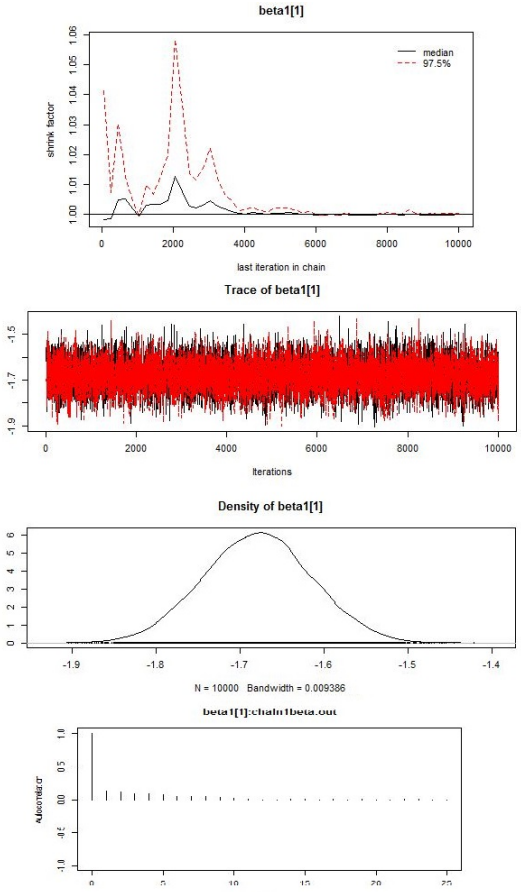

Figure 4.2: Convergence Plot for the Sample Item

4.2 Results of Simulation Study 1

The results of the first simulation study are presented in this section. As mentioned earlier, each data set was generated to have one class. The data generated by the RM were fitted with the MRM, and the data that were generated by 2PL and 3PL IRT models were fitted with Mix2PL and Mix3PL IRT models, respectively. These three models were fit with a one-class

solution and a two-class solution using standard normal priors on ability parameters for each simulation condition. As stated above, five distributions (i.e., bimodal symmetric, normal, platykurtic, skewed, and uniform), two test lengths (10 and 28), two sample sizes (600 and 2,000) and three MixIRT models were manipulated in the first simulation study. In the following parts, we first present the recovery of item parameter estimates and then present the results of correct classification rates under different conditions of the first simulation study.

4.2.1 Recovery of Item parameters

As previously indicated, the recovery of item parameters was assessed using RMSE and bias between the generating parameters and the parameter estimates of one-class MixIRT models. Item parameter estimates from one-class solution MixIRT analyses were used for recovery analyses because the one-class solution was the MixIRT equivalent of the single-class IRT models that were used for data generation. The estimated parameters of one-class MixIRT models were placed on the scale of the generating parameters using the mean/mean transformation method in order to achieve an appropriate common metric. Mean RMSE and mean bias values for each condition were tabulated for item difficulty, item discrimination, and item guessing parameters, separately.

In addition to reporting mean RMSE and mean bias values, the recovery of item parameters were examined using analysis of variance (ANOVA) technique. As is the case with most parametric statistical models, ANOVA is based on several assumptions including normality, homogeneity of variance and independence of observations (Howell, 2014). Data should meet these assumptions in order to make correct statistical inferences based on F distribution. In ANOVA, we assume normality and constant variance for the model error term. In addition, we assume that the residual effects (i.e., error terms) are independent from observation to observation. Although, ANOVA is robust to violations of the normality and homogeneity of

variance assumptions, it is not robust to violation of the independence assumption. Thus, it is important to ensure that data fit these assumptions before conducting analyses. In the case of violations of these assumptions, the results from analyses of the raw data may be untrustworthy. To overcome this problem, one can perform standard statistical tests with either transformed response variables or nonparametric techniques. In this study, the normality assumption was tested by examining the Q-Q plot of residual values, and the equal variance assumption was tested by using Bartlett's test.

For each item parameter, two separate factorial ANOVAs were conducted using RMSE and bias values as dependent variables. Factor variables were the number of items (k), number of examinees (N), distribution condition (D), and the model type (M). All main and interaction effects were included in the ANOVAs. Prior to conducting each ANOVA analysis, the normality and equal variance assumptions were tested and transformed dependent variables were used when necessary. Sum of squares, mean squares, values of effect size and F test with p-value were tabulated for each item parameter. Eta-squared (η^2) was used as the effect size in this study due to easy interpretation (Levine & Hullett, 2002). Eta-squared permits quantifying the effectiveness of a particular intervention (e.g., sample size) by providing information about the magnitude of the effect. Eta-squared is simply the ratio of sum of squares of treatment to total sum of squares and ranges from 0 to 1. Thus, we can interpret eta-squared as the percent of variance accounted for by a factor. General rules of thumb given by Cohen (1988) suggest that eta-squared values that are greater than .14 are considered large effects. This criterion was used in this study.

Recovery of Item Difficulty Parameters. Item difficulty parameter estimate results are presented first. This parameter is the common item parameter for all three MixIRT models considered in this study. Mean RMSE and mean bias values for item difficulty parameters of three MixIRT models for each condition are presented in Tables 4.1 and 4.2. Condition names are given in the first column of Tables 4.1 and 4.2 and include model name, number

of items, and number of examinees. For example, the condition RM282000 indicates a data condition that was generated with the Rasch model for 28 items and 2,000 examinees.

Table 4.1: Mean RMSE Values of Item Difficulty Parameters over 50 Replications

Condition	Bimodal	Normal	Platykurtic	Skewed	Uniform
RM10600	0.167	0.091	0.089	0.091	0.098
RM28600	0.140	0.093	0.094	0.094	0.095
RM102000	0.147	0.067	0.071	0.064	0.078
RM282000	0.097	0.051	0.050	0.052	0.057
2PL10600	0.340	0.197	0.192	0.182	0.195
2PL28600	0.279	0.131	0.133	0.136	0.134
2PL102000	0.362	0.108	0.124	0.113	0.153
2PL282000	0.289	0.070	0.071	0.077	0.105
3PL10600	0.591	0.301	0.291	0.311	0.346
3PL28600	0.560	0.230	0.220	0.267	0.260
3PL102000	0.606	0.256	0.256	0.282	0.282
3PL282000	0.568	0.154	0.163	0.276	0.268

Table 4.1 summarizes the mean RMSE values of item difficulty parameters for three MixIRT models. Mean RMSE values of item difficulty parameter for MRMs were found to be less than 0.10 for most of the conditions. The mean RMSE values were around 0.15 in only three of the bimodal data conditions. As shown in Table 4.1, mean RMSE values of the Mix2PL and Mix3PL IRT models were larger than those for the MRM. A pattern

of increasing RMSE values appears to be present in which RMSE values increase as the complexity of the model increases. The mean RMSEs for the Mix2PL IRT model condition with 28 items and 2,000 examinees, however, were less than 0.11 for all except the bimodal symmetric distribution. For the Mix2PL analyses, mean RMSE values seemed to decrease as the number of examinees and the number of items increased. The mean RMSE values for the bimodal distribution were relatively higher for the Mix3PL IRT model. Mean RMSE values were around 0.30 for normal, platykurtic, skewed, and uniform distributions. The Mix3PL condition with 28 items and 2,000 examinees yielded the smallest mean RMSE values for all except the bimodal symmetric distribution. These results are consistent with previous simulation studies with MixIRT models (Li et al., 2009).

Table 4.2 summarizes the mean bias values of item difficulty parameters for three MixIRT models. The absolute values of mean bias values for item difficulty parameters were found to be less than 0.05 for all of the MRM conditions. The best performing conditions for MRMs were the conditions with platykurtic distribution as they produced the smallest bias values across all simulation conditions. MRM conditions with normal and skewed distributions displayed similar performances to platykurtic distribution conditions. Bimodal symmetric distribution condition yielded the largest mean bias values for the MRMs. Conditions with uniform distributions were the second largest biased conditions for recovering item difficulty parameters of MRMs. The mean bias values for item difficulty parameters tended to increase as the model complexity increased. Mean bias values of item difficulty parameters for Mix2PL IRT model analyses were found to be less than 0.05 for more than half of the 2PL IRT data conditions. The mean bias values were around 0.12 in only three of the bimodal data conditions. Consistent with the bias results of MRMs, conditions with bimodal distributions were more biased than the conditions with other distributions. The normal, platykurtic, skewed, and uniform distribution conditions showed similar performances in terms of bias for item difficulty parameters of the Mix2PL IRT model analyses. Overall,

Mix3PL conditions were the most biased conditions for all of the distributions. Mean bias values for the item difficulty parameter were low and generally in the negative direction for the conditions with normal, platykurtic, and skewed distributions. Consistent with previous results reported above, the mean bias values also were the highest for the bimodal symmetric data conditions in Mix3PL analyses.

Table 4.2: Mean Bias Values of Item Difficulty Parameters over 50 Replications

Condition	Bimodal	Normal	Platykurtic	Skewed	Uniform
RM10600	−0.043	−0.001	−0.001	−0.004	−0.016
RM28600	0.031	0.011	0.007	0.003	−0.012
RM102000	−0.049	−0.003	−0.002	−0.004	−0.022
RM282000	−0.003	0.004	−0.002	−0.005	−0.015
2PL10600	0.044	0.086	0.102	0.067	0.086
2PL28600	0.121	0.029	0.034	0.024	0.035
2PL102000	0.103	0.045	0.048	0.051	0.038
2PL282000	0.120	0.006	0.007	0.007	0.010
3PL10600	0.548	−0.108	−0.065	−0.083	0.171
3PL28600	0.516	−0.058	−0.059	−0.158	0.103
3PL102000	0.587	−0.091	−0.026	−0.171	0.145
3PL282000	0.530	−0.038	−0.014	−0.230	0.171

ANOVA Results for Examining the Recovery of Item Difficulty Parameters.

As indicated above, the normality and equal variance assumptions were tested prior to conducting ANOVA analyses. Since raw RMSE and bias values for the item difficulty parameter violated these assumptions, the natural logarithm of RMSE (log[RMSE]) and bias (log[bias]) values were used in the ANOVA analyses. The ANOVA table for log[RMSE] of item difficulty is presented in Table 4.3. Similarly, Table 4.4 summarizes the ANOVA results for log[bias] values. Results in both of the tables indicate that model complexity is an important factor for recovery of item difficulty parameter estimates due to large and significant eta-squared values. Thus, we concluded that as the model complexity increased (e.g., from MRM to Mix2PL IRT model), the difficulty of recovering the item difficulty parameter estimates increased. This was also shown in the Tables of RMSE and bias for item difficulty parameter estimates.

In addition, the ANOVA table for log[bias] showed that the distribution condition also had a significantly large effect size ($\eta^2 = .378$). Distribution condition was also significant in the ANOVA analysis of log[RMSE] but with a moderate (i.e., .125) eta-squared value. As shown in Table 4.3, a number of factors (i.e., *k*, *N*, *N*D*, *N*M*, and *D*M*) were found to be significant with small effect sizes ($\eta^2 < 0.054$). As shown in the ANOVA table of log[bias] variable, a number of factors (*k*, *N*, *N*D*, and *D*M*) were significant with small effect size values in addition to the model complexity and distribution conditions that were mentioned above. Based on the results of these two ANOVAs, we can conclude that there are likely meaningful differences in the RMSE and bias among the three MixIRT models. It appears that there is also meaningful difference in the bias among the five distributions.

Table 4.3: ANOVA Results for log[RMSE] of Item Difficulty Parameter Estimates

Source	*df*	Sum of Square	Mean Square	F	p	η^2
k	1	9.702	9.702	44.720	<.001	.015
N	1	35.154	35.154	162.050	<.001	.054
$k*N$	1	0.780	0.780	3.590	.058	.001
D	4	80.851	20.213	93.170	<.001	.125
$k*D$	4	0.919	0.230	1.060	.376	.001
$N*D$	4	5.258	1.315	6.060	<.001	.008
$k*N*D$	4	0.335	0.084	0.390	.819	.001
M	2	266.982	133.491	615.340	<.001	**.411**
$k*M$	2	0.752	0.376	1.730	.177	.001
$N*M$	2	5.360	2.680	12.350	<.001	.008
$k*N*M$	2	0.785	0.392	1.810	.164	.001
$D*M$	8	4.502	0.563	2.590	.008	.007
$k*D*M$	8	0.966	0.121	0.560	.814	.002
$N*D*M$	8	1.896	0.237	1.090	.366	.003
$k*N*D*M$	8	0.350	0.044	0.200	.991	.001

Note. k = number of items, N = number of examinees, D = distribution condition, and M = model type.

Table 4.4: ANOVA Results for log[bias] of Item Difficulty Parameter Estimates

Source	*df*	Sum of Square	Mean Square	*F*	*p*	η^2
k	1	38.826	38.826	32.560	<.001	.020
N	1	5.231	5.231	4.390	.037	.003
*k*N*	1	0.888	0.888	0.740	.389	.001
D	4	735.766	183.942	154.250	<.001	**.378**
*k*D*	4	11.174	2.794	2.340	.054	.006
*N*D*	4	11.918	2.979	2.500	.042	.006
*k*N*D*	4	0.417	0.104	0.090	.986	.000
M	2	408.226	204.113	171.170	<.001	**.210**
*k*M*	2	2.668	1.334	1.120	.328	.001
*N*M*	2	4.159	2.079	1.740	.176	.002
*k*N*M*	2	1.738	0.869	0.730	.483	.001
*D*M*	8	23.056	2.882	2.420	.014	.012
*k*D*M*	8	10.050	1.256	1.050	.395	.005
*N*D*M*	8	12.965	1.621	1.360	.212	.007
*k*N*D*M*	8	3.198	0.400	0.340	.952	.002

Note. k = number of items, N = number of examinees, D = distribution condition, and M = model type.

Recovery of Item Discrimination Parameters. Results are presented here for item discrimination parameter for Mix2PL and Mix3PL models. Mean RMSE and mean bias values for item discrimination parameters for each condition are presented in Tables 4.5 and 4.6. Again, condition names that are given in the first column of Tables 4.5 and 4.6 include model name, number of items, and number of examinees. Conditions were labeled in these tables as was done in previous tables. So, the condition 2PL282000 indicates a data condition generated with the 2PL IRT model for 28 items and 2,000 examinees and analyzed with the Mix2PL IRT model. Mean RMSE values for the item discrimination parameter estimates for the Mix2PL and Mix3PL IRT models are presented in Table 4.5. As expected, RMSE values for the Mix2PL and Mix3PL IRT models for the bimodal symmetric distribution were the largest. The RMSE values for the uniform distribution were the second largest. Mean RMSE values appeared to be smaller for all of the Mix2PL conditions under the normal, platykurtic, and skewed distributions. Mean RMSE values of item discrimination parameter seemed to decrease as the number of examinees and the number of items increased for most of the conditions. Mean RMSE values for the Mix3PL IRT model conditions yielded higher values than those for Mix2PL IRT analyzes. Consistent with the results of Mix2PL IRT model analyses, bimodal symmetric and uniform distribution conditions yielded the largest RMSE values for item discrimination parameter estimates of the Mix3PL IRT model.

Table 4.5: Mean RMSE Values of Item Discrimination Parameters over 50 Replications

Condition	Bimodal	Normal	Platykurtic	Skewed	Uniform
2PL10600	1.684	0.150	0.143	0.154	0.301
2PL28600	1.528	0.129	0.130	0.140	0.285
2PL102000	1.786	0.086	0.090	0.089	0.358
2PL282000	1.801	0.071	0.071	0.080	0.369
3PL10600	1.457	0.272	0.291	0.409	0.431
3PL28600	1.497	0.270	0.278	0.379	0.423
3PL102000	1.581	0.231	0.233	0.499	0.426
3PL282000	2.022	0.204	0.214	0.598	0.440

Table 4.6 summarizes the mean bias values of item discrimination parameters for the Mix2PL and Mix3PL IRT models. Mean bias values for item discrimination parameter for Mix2PL IRT model analyses were found to be less than 0.025 for the normal, platykurtic, and skewed distribution conditions. The lowest performing conditions for Mix2PL IRT models were the conditions with bimodal symmetric distributions. These are the ones that produced the highest bias values across all simulation conditions. The Mix2PL conditions with uniform distributions yielded the second-most biased estimates for item discrimination parameters. Consistent with previous results (above) for the Mix2PL IRT model analyses, the Mix3PL IRT data conditions with bimodal distributions were the most biased conditions, when analyzed with the Mix3PL IRT model. Uniform distribution conditions yielded higher mean bias values than the normal, platykurtic, and skewed distribution conditions. There was no

clear pattern, however, with respect to effects of the number of items and sample size on bias calculation for item discrimination parameter.

Table 4.6: Mean Bias Values of Item Discrimination Parameters over 50 Replications

Condition	Bimodal	Normal	Platykurtic	Skewed	Uniform
2PL10600	−0.835	−0.003	−0.002	−0.001	−0.132
2PL28600	−0.759	0.023	0.022	0.024	−0.128
2PL102000	−0.890	−0.002	−0.005	−0.001	−0.172
2PL282000	−0.898	0.006	0.002	0.009	−0.181
3PL10600	−0.460	0.223	0.219	0.079	0.101
3PL28600	−0.736	0.020	0.023	−0.131	−0.156
3PL102000	−0.785	−0.045	−0.046	−0.222	−0.193
3PL282000	−1.000	−0.010	−0.001	−0.276	−0.190

ANOVA Results for Recovery of Item Discrimination Parameters. As with the previous ANOVA analyses, the normality and equal variance assumptions were tested prior to performing ANOVAs with RMSE and bias for the item discrimination parameter. Since raw RMSE and bias values for the item discrimination parameter violated these assumptions, the natural logarithm of RMSE (log[RMSE]) and square root of bias (sqrt[bias]) values were used in the ANOVA analyses. The ANOVA tables for log[RMSE] and sqrt[bias] of item discrimination are presented in Tables 4.7 and 4.8. As shown in Table 4.7, a number of the main and interaction effects of the conditions of the simulation were significant. Only the distribution shape condition had a large effect ($\eta^2 = .554$), however, for RMSE of item discrimination estimates. The model type had a moderate effect size and the other significant

factors (k, k*D, k*M, N, N*D, N*M, D*M, and N*D*M) had small effects ($\eta^2 < .05$) on the RMSE calculation for item discrimination parameter. As shown in Table 4.8, only model type (Mix2PL vs. Mix3PL) had a large and significant effect on bias calculation for the item discrimination parameter. In addition, the distribution condition had a moderate and significant effect, but the sample size and interaction of distribution and model type had small and significant effects on bias calculation for the item discrimination parameter. Thus, we can conclude that there are likely meaningful differences in both of the RMSE and bias among the two MixIRT models and the five distribution conditions.

Table 4.7: ANOVA Results for RMSE

of Item Discrimination Parameter Estimates

Source	*df*	Sum of Square	Mean Square	*F*	*p*	η^2
k	1	0.425	0.425	7.620	.006	.001
N	1	2.365	2.365	42.460	<.001	.003
*k*N*	1	0.036	0.036	0.640	.424	.000
D	4	450.909	112.727	2023.820	<.001	**.554**
*k*D*	4	0.748	0.187	3.360	.010	.001
*N*D*	4	8.257	2.064	37.060	<.001	.010
*k*N*D*	4	0.375	0.094	1.680	.152	.001
M	1	65.031	65.031	1167.520	<.001	.080
*k*M*	1	0.487	0.487	8.750	.003	.001
*N*M*	1	3.113	3.113	55.890	<.001	.004
*k*N*M*	1	0.040	0.040	0.710	.400	.000
*D*M*	4	39.614	9.904	177.800	<.001	.049
*k*D*M*	4	0.155	0.039	0.690	.596	.000
*N*D*M*	4	4.952	1.238	22.220	<.001	.006
*k*N*D*M*	4	0.169	0.042	0.760	.552	.000

Note. k = number of items, N = number of examinees, D = distribution condition, and M = model type.

Table 4.8: ANOVA Results for Bias

of Item Discrimination Parameter Estimates

Source	*df*	Sum of Square	Mean Square	*F*	*p*	η^2
k	1	0.009	0.009	0.940	.333	.002
N	1	0.184	0.184	19.640	<.001	.041
*k*N*	1	0.011	0.011	1.180	.279	.002
D	4	0.281	0.094	9.980	<.001	.062
*k*D*	4	0.010	0.003	0.340	.794	.002
*N*D*	4	0.042	0.014	1.490	.219	.009
*k*N*D*	4	0.028	0.009	0.990	.398	.006
M	1	1.747	1.747	185.960	<.001	**.383**
*k*M*	1	0.016	0.016	1.710	.193	.004
*N*M*	1	0.002	0.002	0.180	.674	.000
*k*N*M*	1	0.026	0.026	2.760	.098	.006
*D*M*	4	0.148	0.074	7.900	.001	.033
*k*D*M*	4	0.009	0.004	0.450	.636	.002
*N*D*M*	4	0.001	0.001	0.120	.724	.000
*k*N*D*M*	4	0.008	0.008	0.840	.360	.002

Note. k = number of items, N = number of examinees, D = distribution condition, and M = model type.

Recovery of Item Guessing Parameters. Table 4.9 presents mean RMSE values of the item guessing parameter estimates for each condition. Table 4.10 summarizes mean bias values of the item guessing parameter. Condition names that are given in the first column of Tables 4.9 and 4.10 was labeled as in previous tables. As can be seen in Table 4.9, the skewed data conditions yielded the highest mean RMSE values for the item guessing parameter. For most of the conditions with 28 items, mean RMSE values for the item guessing parameter appeared to decrease as the number of examinees increased. However, no clear pattern emerged with regard to changes in the number of items. The mean RMSE values for item guessing parameters were relatively lower than those for item difficulty and discrimination parameters. This is because the item guessing parameter estimates are between zero and one whereas estimates for discrimination and difficulty parameters can take on larger absolute values.

Table 4.9: Mean RMSE Values of Item Guessing Parameters over 50 Replications

Condition	Bimodal	Normal	Platykurtic	Skewed	Uniform
3PL10600	0.059	0.066	0.065	0.071	0.063
3PL28600	0.050	0.068	0.066	0.088	0.058
3PL102000	0.051	0.069	0.069	0.090	0.059
3PL282000	0.039	0.054	0.054	0.113	0.055

Table 4.10 summarizes the mean bias values of item guessing parameters for the Mix3PL IRT models. Consistent with the RMSE statistics, mean bias values in the skewed data conditions were higher than those for the other distributions. The mean bias values for item guessing parameters were generally in the negative direction for the normal, platykurtic, skewed, and most of the uniform data conditions. For the item guessing parameter estimates

for the Mix3PL IRT model analyses, the bimodal symmetric conditions yielded lower mean bias values than those for item difficulty and discrimination parameters of previous MixIRT analyses. For the bimodal symmetric distribution conditions, bias was consistently positive, which indicated that the guessing parameter was overestimated.

Table 4.10: Mean Bias Values of Item Guessing Parameters over 50 Replications

Condition	Bimodal	Normal	Platykurtic	Skewed	Uniform
3PL10600	0.006	−0.016	−0.010	−0.020	−0.003
3PL28600	0.006	−0.018	−0.014	−0.035	−0.006
3PL102000	0.009	−0.013	−0.007	−0.036	0.001
3PL282000	0.006	−0.013	−0.008	−0.053	−0.001

ANOVA Results for Recovery of Item Guessing Parameters. Two separate ANOVAs were performed using the natural logarithm of RMSE (log[RMSE]) and the square root of bias (sqrt[bias]) as dependent variables since the raw RMSE and bias for the item guessing parameter violated the assumptions of ANOVA. The only difference between ANOVAs for the item guessing parameter and previous ANOVAs is the lack of the model factor. This was because only the Mix3PL IRT model had the item guessing parameter. The ANOVA tables for log[RMSE] and sqrt[bias] of item guessing are presented in Tables 4.11 and 4.12. As shown in Table 4.11, the distribution condition had a significant and a moderate ($\eta^2 = .121$) effect while the interaction of distribution and sample size (i.e., $N*D$) had a small ($\eta^2 = .023$) and significant effect on the RMSE calculation for item guessing parameter. As shown in Table 4.12, the number of items (k) had a moderate ($\eta^2 = .074$) and significant effect while the sample size (N) and distribution (D) conditions had small and significant effects on the bias calculation for item guessing parameter estimates.

Table 4.11: ANOVA Results for RMSE
of Item Guessing Parameter Estimates

Source	*df*	Sum of Square	Mean Square	*F*	*p*	η^2
k	1	0.295	0.295	1.000	.318	.002
N	1	0.061	0.061	0.210	.650	.001
*k*N*	1	0.312	0.312	1.060	.305	.003
D	4	15.369	3.842	13.020	<.001	.121
*k*D*	4	1.494	0.374	1.270	.283	.012
*N*D*	4	2.875	0.719	2.440	.047	.023
*k*N*D*	4	0.378	0.094	0.320	.865	.003

Note. k = number of items, N = number of examinees, and D = distribution condition.

Table 4.12: ANOVA Results for Bias

of Item Guessing Parameter Estimates

Source	*df*	Sum of Square	Mean Square	*F*	*p*	η^2
k	1	0.060	0.060	16.040	<.001	.074
N	1	0.044	0.044	11.860	.001	.054
*k*N*	1	0.000	0.000	0.130	.716	.001
D	4	0.046	0.012	3.110	.017	.057
*k*D*	4	0.002	0.001	0.140	.969	.003
*N*D*	4	0.012	0.003	0.800	.526	.015
*k*N*D*	4	0.008	0.002	0.520	.725	.009

Note. k = number of items, N = number of examinees, and D = distribution condition.

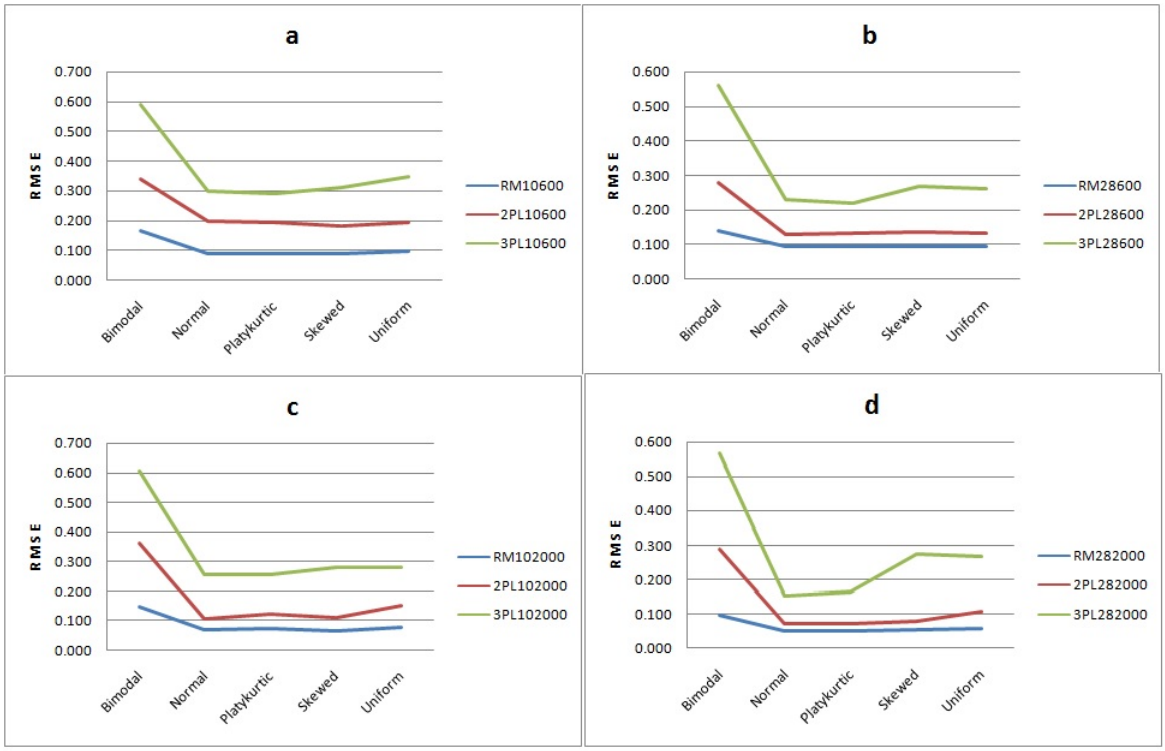

Figure 4.3: RMSE Plots for Item Difficulty Parameter Estimates

Summary of the Recovery Assessment of Simulation Study 1. Two evaluative criteria were used for the item parameter estimates: the mean RMSE and bias of the estimates. For the difficulty parameter, the MRMs tended to produce the smallest RMSE and bias values. The RMSE results for item difficulty parameter estimates are illustrated for each data condition in Figure 4.3. As can be seen in Figure 4.3, the accuracy of the item difficulty parameter appeared to decrease as the model complexity increased. The distribution condition also appeared to have an important effect on the recovery of the item difficulty parameter (see Figure 4.3). It was more difficult to recover item difficulty parameters, particularly, for the bimodal symmetric and uniform conditions. Compared to the effects of the model complexity and distribution factors, the number of examinees and items seemed to be relatively less important factors on the recovery of item difficulty parameter estimates.

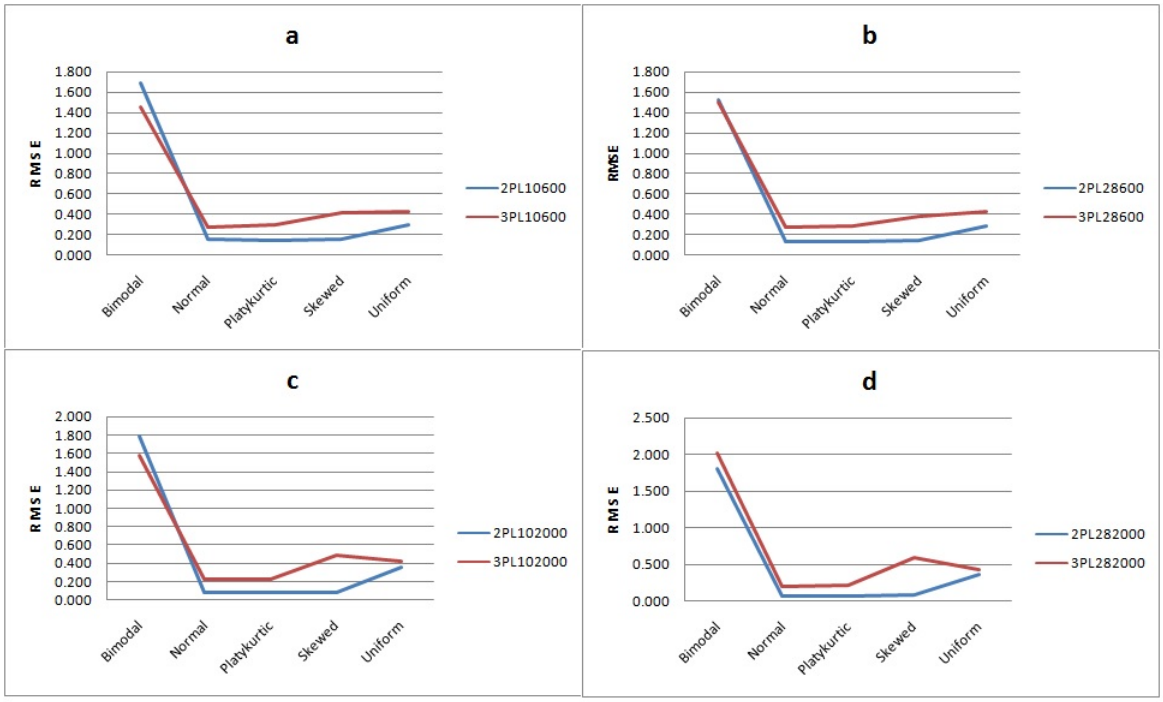

Figure 4.4: RMSE Plot for Item Discrimination Parameter Estimates

The RMSE results for item discrimination parameter estimates are illustrated for the Mix2PL and Mix3PL IRT models in Figure 4.4. For the discrimination parameter, conditions with bimodal symmetric distributions tended to produce more biased estimates than those with other distributions. The conditions with uniform distributions appeared to have the second most biased estimates. ANOVA analyses also provide evidence for the significance of effects of distribution type on the recovery of item discrimination parameter estimates. The item discrimination parameter estimates from Mix3PL model analyses appeared to be more biased than those from the Mix2PL IRT model analyses. The number of examinees and items did not appear to affect recovery of the item discrimination parameter.

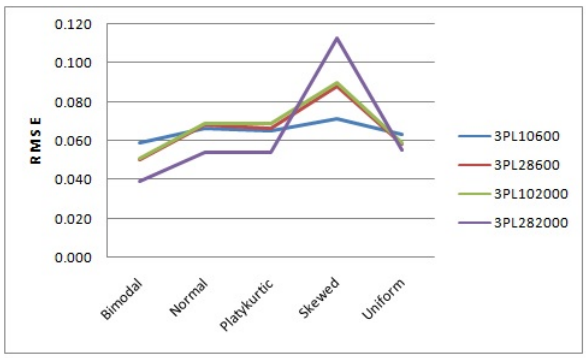

Figure 4.5: RMSE Plot for Item Guessing Parameter Estimates

The RMSE results for item guessing parameter estimates are illustrated for the Mix3PL IRT model in Figure 4.5. Almost all of the conditions yielded very small RMSE (<0.10) and bias (<0.020) values for item guessing parameter estimates. Mix3PL model analyses appeared to have difficulty producing accurate item difficulty and discrimination estimates. ANOVA analyses with RMSE and bias values showed that distribution type had an important effect on the recovery of the item guessing parameter.

4.2.2 Correct Positive Rates

The proportions of correct positive detection for the three MixIRT models were calculated based on minimum AIC and BIC between one-class and two-class solutions. For instance, the number of classes for the given data set was defined as correct, when the considered information index for a one-class solution was smaller than that for the two-class solution. These proportions are presented in Tables 4.13, 4.14, and 4.15 for each condition of the MRM, Mix2PL, and Mix3PL IRT models, respectively. Condition names that are given in the first column of Tables 4.13 to 4.15 include model name, number of items and number of examinees. For example, the condition RM102000 indicates a data condition that was generated with the Rasch model for 10 items and 2,000 examinees.

Table 4.13: The Correct Positive Rates for MRM Analyses over 50 Replications

Condition	Bimodal		Normal		Platykurtic		Skewed		Uniform	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
RM10600	0	0	6	49	10	47	3	41	0	3
RM102000	0	0	0	39	0	40	0	22	0	0
RM28600	0	0	42	50	41	50	33	50	19	49
RM282000	0	0	35	50	31	50	1	50	0	0

Table 4.13 summarizes the correct positive rates for MRM analyses. As shown in this table, the BIC index performed well in the MRM analysis for the normal, platykurtic, and skewed conditions. The proportions of correct positives for the BIC index for the bimodal symmetric and uniform conditions, however, were low except for the uniform distribution condition with 28 items and 600 examinees. The proportions of correct positives for the bimodal symmetric distribution was zero for all of the conditions. The performance of the BIC index appeared to increase as the number of examinees decreased and as the number of items increased. The overall performance of AIC was lower than BIC for the MRM analyses. The AIC did not provide high correct identification rates in the normal distribution conditions, especially with 10 items. Similarly, for the platykurtic and skewed distribution conditions, the AIC index showed poor performance for 10-item conditions. The performance of AIC index was higher for 28-item conditions, however, than those for 10-item data conditions that were generated under normal, platykurtic, and skewed distributions. The proportions of correct positives for AIC were zero for all of the bimodal symmetric distribution conditions. The second lowest performance of the AIC index was observed with the results of uniform

distribution except for the condition with 28 items and 600 examinees. In that condition, 19 out of the 50 replications were correct detections. The performance of the AIC index appeared to increase as the number of items increased. However, the proportions of correct positives for the AIC index were lower in large sample size conditions. Both AIC and BIC showed good performance in data conditions with 28 items and 600 examinees except in the bimodal data condition. Based on the results of the AIC and BIC indices, latent non-normality seemed to be a problem for the data conditions with the bimodal symmetric and uniform distributions.

Table 4.14: The Correct Positive Rates for Mix2PL IRT Model Analyses over 50 Replications

Condition	Bimodal		Normal		Platykurtic		Skewed		Uniform	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
2PL10600	12	50	16	46	15	48	1	29	15	50
2PL102000	18	49	2	40	7	46	0	6	2	32
2PL28600	48	50	49	50	47	50	20	49	34	50
2PL282000	34	50	47	50	45	50	0	24	2	50

Table 4.14 presents the correct positive rates for Mix2PL IRT model analyses. For almost all conditions, the correct positive rates of the BIC index were found to be close to 100 percent. This was not the case for the skewed data conditions. For the skewed data conditions, the BIC index showed good performance (49 out of 50 correct detections) for the 28 items and 600 examinees condition. The BIC index showed 100% performance for all of the 28-item conditions except for the skewed distribution. For the 10-item conditions, the performance of the BIC index decreased as the number of examinees increased. The results of the AIC index in the Mix2PL IRT model analyses provided higher correct detection rates than those

of the MRM analyses. Overall performance of AIC index was worse than the BIC results in Mix2PL IRT model analyses. Correct positive selection rates for AIC ranged from 0 to 20 percent in more than half of the conditions. Consistent with the results for the BIC index, the performance of the AIC index was generally lower in the 10-item conditions than in the 28-item conditions. Further, the performance of the AIC index decreased as the number of examinees increased. The AIC index showed poor performance for the data conditions with skewed and uniform distributions. Based on the results from AIC index, latent non-normality appears to cause detection of spurious latent class in the Mix2PL model. Results based on the BIC index, however, were more accurate for this model in these conditions.

Table 4.15: The Correct Positive Rates for Mix3PL Analyses over 50 Replications

Condition	Bimodal		Normal		Platykurtic		Skewed		Uniform	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
3PL10600	46	50	10	50	9	50	14	50	7	48
3PL102000	18	50	1	36	2	33	3	46	0	44
3PL28600	50	50	49	50	48	50	50	50	48	50
3PL282000	46	50	42	50	14	50	17	50	7	50

Table 4.15 presents the correct positive detection rates for Mix3PL model analyses. In all distribution conditions, BIC supported selection of the correct one class model in 100 % of the replications except for the 10 items and 2,000 examinees conditions. The conditions with the normal and skewed distributions yielded the lowest BIC detection results. The number of correct selections was higher for the AIC for the Mix3PL model in comparison to the other mixture IRT models. Consistent with previous results, however, the number of correct model

selections by AIC was lower than for BIC. Further, AIC was less accurate in selecting the correct model in all of the uniform data conditions except for the 28 items and 600 examinees conditions. The AIC generally was less effective at detection of the correct model for the 10 items and 2,000 examinees one-class condition as well. The performances of both AIC and BIC indices appeared to improve as the number of items increased and as the number of examinees decreased. These results suggest that the Mix3PL models may be more robust to latent non-normality than either the MRM or Mix2PL IRT models.

Summary of the Correct Positive Detection Rates of Simulation Study 1. According to the results of the AIC index, latent non-normality appeared to be a problem for MRM analyses, when using a normal prior for ability parameters. In addition, the BIC index showed that MRM analyses of the data with bimodal symmetric and uniform distributions tended to produce spurious latent classes. The performance of the BIC index, however, was better than the AIC index in determining the correct number of latent classes in the MRM analyses. The performances of both indices appeared to increase as the number of items increased and as the number of examinees decreased. For the Mix2PL IRT model analyses, the correct positive rates of the BIC index were close to 100% for most of the conditions. Results for the AIC index suggested that non-normality caused spurious latent class detections for the Mix2PL model. Analyses for this model with 28 items showed fewer over-extraction problems, however, than for the 10-item conditions. As with the MRMs, the correct positive rates for Mix2PL model analyses increased as the number of examinees decreased. A pattern of increasing correct positive rates appeared to be present in which correct positive rates increased as the complexity of the model increased. The Mix3PL IRT model analyses resulted in over-extraction of latent classes in a few conditions based on the BIC index. However, the performance of the AIC index showed that spurious latent class was also a concern for non-normally distributed data conditions, especially for the 10-item

conditions. AIC model selection for the Mix3PL analyses with 28 items yielded higher correct positive rates than the 10-item data conditions.

4.3 Results of Simulation Study 2

The results of the second simulation study are presented in this section. As with the results for Simulation Study 1, each data set was generated to have one class. The data that were generated by the RM were fitted with the MRM, and the data that were generated by 2PL and 3PL models were fitted with the Mix2PL and Mix3PL IRT models, respectively. These three models were fit with a one-class solution and a two-class solution using standard normal priors on ability parameters for each simulation condition. As stated above, six combinations of skewness and kurtosis values were used to generate simulated data in this second simulation study to investigate the effects of these distributional parameters on detection of spurious latent classes. Six distributions, two test lengths (10 and 28), two sample sizes (600 and 2,000), and three MixIRT models were manipulated in the second simulation study. The first three distribution conditions were generated to have the same skewness value (i.e., 1.5) and different kurtosis values while the last three conditions were generated to have the same kurtosis values (i.e., 3.5) and different skewness values. The kurtosis values were 4.0, 3.50, and 2.50 for Conditions 1, 2, and 3, respectively. The skewness values were 1.0, 0.50, and 0.0 for Conditions 4, 5, and 6, respectively (see also Table 4.16). Below, the recovery of item parameter estimates are presented followed by the correct classification rates under the different conditions.

Table 4.16: Six Data Conditions for Simulation Study 2

Condition	Skewness	Kurtosis
Condition 1	1.50	4.00
Condition 2	1.50	3.50
Condition 3	1.50	2.50
Condition 4	1.00	3.50
Condition 5	0.50	3.50
Condition 6	0.00	3.50

4.3.1 Recovery of Item parameters

As indicated above, the item parameter estimates from one-class solution MixIRT analyses were used to calculate RMSE and bias values for recovery assessments. The estimated parameters of one-class MixIRT models were placed on the scale of the generating parameters using the mean/mean transformation method. Mean RMSE and bias values for each condition were tabulated for item difficulty, item discrimination, and item guessing parameters, separately. The recovery of item parameters was also examined using ANOVA analyses. As with the first simulation study, the normality assumption was tested by examining Q-Q plot of residual values, and equal variance assumption was tested using Bartlett's test. ANOVA analyses were conducted using transformed RMSE and bias values when necessary.

Recovery of Item Difficulty Parameters. Item difficulty parameter results for the second simulation study are discussed first as this parameter is common to all three MixIRT models. The mean RMSE and mean bias values for the item difficulty parameter for each condition are presented in Tables 4.17 and 4.18. The condition names that are given in the first column of Tables 4.17 and 4.18 include model name, number of items, and number of examinees. For example, the condition 3PL10600 indicates a data condition that was generated with the 3PL model for 10 items and 600 examinees.

Table 4.17: Mean RMSE Values of Item Difficulty Parameters over 50 Replications

Condition	COND1	COND2	COND3	COND4	COND5	COND6
RM10600	0.146	0.108	0.097	0.102	0.105	0.104
RM28600	0.132	0.095	0.092	0.098	0.096	0.095
RM102000	0.103	0.065	0.066	0.059	0.059	0.062
RM282000	0.086	0.057	0.056	0.057	0.055	0.058
2PL10600	0.251	0.206	0.205	0.188	0.196	0.206
2PL28600	0.178	0.138	0.136	0.137	0.134	0.143
2PL102000	0.177	0.129	0.127	0.123	0.131	0.121
2PL282000	0.121	0.100	0.099	0.092	0.087	0.092
3PL10600	0.350	0.368	0.325	0.373	0.352	0.354
3PL28600	0.334	0.344	0.370	0.314	0.314	0.261
3PL102000	0.392	0.399	0.382	0.388	0.358	0.323
3PL282000	0.437	0.432	0.490	0.329	0.258	0.214

As shown in Table 4.17, mean RMSE values for the item difficulty parameter of MRMs were found to be less than 0.10 for most of the conditions. The mean RMSE values were below 0.15 for the rest of the MRM conditions. The mean RMSE values appeared to decrease as the kurtosis values decreased from Condition 1 to Condition 3. However, no clear pattern emerged for the mean RMSE of item difficulty parameter with regard to change of skewness values across the last three conditions. Overall, the mean RMSE of item difficulty estimates indicated good recovery for the MRMs. In Table 4.17, the mean RMSEs for item difficulty parameters in the 2PL data conditions ranged from 0.087 to 0.251. Mean RMSEs of item difficulty parameter tended to decrease as the number of examinees and items increased for the MRMs and Mix2PL IRT analyses. Additionally, the mean RMSE values tended to decrease as the kurtosis values decreased in the 2PL data conditions. Similar to the MRM results, skewness values displayed no clear pattern in terms of mean RMSE values in the MiX2PL analyses. A pattern of increasing RMSE values appeared to be present in which RMSE values increased as the complexity of the model increased. The mean RMSE values for the Mix3PL analyses were larger than those for the Mix2PL model and MRM analyses. Unlike the results of the mean RMSE for the MRMs and Mix2PL model, kurtosis showed no clear pattern in the Mix3PL IRT model analyses. The mean RMSE for the item difficulty parameter estimates, however, appeared to decrease as the skewness values decreased for Condition 4 to Condition 6. Further, the mean RMSE values for item difficulty parameter estimates tended to decrease as the number of items and sample size increased in these last three conditions for the Mix3PL model analyses.

Table 4.18: Mean Bias Values of Item Difficulty Parameters over 50 Replications

Condition	COND1	COND2	COND3	COND4	COND5	COND6
RM10600	−0.081	−0.025	−0.001	0.006	−0.004	0.008
RM28600	−0.062	0.014	0.009	0.014	0.007	0.015
RM102000	−0.076	−0.021	−0.025	−0.022	−0.011	−0.001
RM282000	−0.052	−0.008	−0.009	−0.017	0.001	0.003
2PL10600	0.102	0.070	0.074	0.064	0.102	0.101
2PL28600	0.008	0.037	0.042	0.033	0.026	0.048
2PL102000	0.039	0.053	0.049	0.045	0.054	0.036
2PL282000	−0.017	0.032	0.038	0.015	0.022	0.026
3PL10600	−0.201	−0.234	−0.191	−0.215	−0.210	−0.208
3PL28600	−0.256	−0.268	−0.300	−0.232	−0.189	−0.167
3PL102000	−0.323	−0.336	−0.316	−0.328	−0.278	−0.230
3PL282000	−0.416	−0.406	−0.467	−0.300	−0.227	−0.164

Results for bias in recovery of item difficulty estimates are shown in Table 4.18. The mean bias values for were less than 0.025 for all conditions except Condition 1. The condition with the most bias for the MRM was Condition 1, which had the highest kurtosis value of 4.0. The mean bias values appeared to decrease as the kurtosis value decreased from Condition 3 to Condition 1 for the Rasch model data. No clear pattern appears to be present for skewness and bias for the MRM item difficulty parameter estimates. The mean bias values for item difficulty parameter of 2PL IRT model data conditions ranged from −0.017 to 0.102. The

mean bias values tended to decrease as the number of items and sample size increased for the 2PL model data conditions. No clear pattern in mean bias emerged for the difficulty parameter estimates for the Mix2PL for either skewness or kurtosis. A pattern of increasing bias values appeared to be present in which mean bias values increased as the complexity of the model increased. Mix3PL model estimates showed the most bias. The absolute mean bias values for item difficulty parameter estimates of Mix3PL model ranged from 0.164 to 0.467. The condition with the smallest skewness value (Condition 6) yielded the least biased estimates for each item and examinee conditions for the Mix2PL model analyses. The item difficulty parameter estimates from first three conditions for Mix3PL analyses appeared to be more biased than those for the last three conditions.

ANOVA Results for Examining the Recovery of Item Difficulty Parameters. Normality and equality of variance assumptions were tested prior to the conducting ANOVA analyses for item difficulty parameter estimates. Since raw RMSE and bias values for item difficulty parameter failed to meet the ANOVA assumptions, the natural logarithm of RMSE (log[RMSE]) and square root of bias (sqrt[bias]) values were used in these ANOVA analyses. The ANOVA tables for log[RMSE] and sqrt[bias] of item difficulty are presented in Tables 4.19 and 4.20. As shown in both tables, the model complexity appeared to be an important factor for recovery of item difficulty parameter estimates due to large and significant eta-squared values. As mentioned above, the factors with larger effect size values were considered to be an important factor. Thus, we concluded that the model complexity had a negative impact on the recovery of item difficulty parameter estimates. That is, it was more difficult to recover item difficulty parameter estimates as the model complexity increased (e.g. from Mix2PL to Mix3PL IRT model).

In addition, a number of factors are shown to have been significant, albeit with small effect sizes, in the ANOVA tables. The number of items (k), sample size (N), distribution condition (D), and four of the two-way interactions of these factors (i.e., N*D, k*M, N*M,

and D^*M) were found to be significant in RMSE calculations. In addition to significant effect of model type, the distribution condition (D) and number of items (k) were significant in bias calculation (see Table 4.20). Based on the results of these two ANOVAs, it appears that there were meaningful differences in the RMSE and bias among the three MixIRT models. Distribution conditions appeared to be the second most important factor for RMSE and bias calculation of item difficulty parameter.

Table 4.19: The Results of ANOVA for RMSE

of Item Difficulty Parameter Estimates

Source	*df*	Sum of Square	Mean Square	*F*	*p*	η^2
k	1	5.310	5.310	41.090	<.001	.008
N	1	28.004	28.004	216.680	<.001	.040
*k*N*	1	0.042	0.042	0.330	.567	.000
D	5	15.485	3.097	23.960	<.001	.022
*k*D*	5	0.569	0.114	0.880	.493	.001
*N*D*	5	1.456	0.291	2.250	.047	.002
*k*N*D*	5	0.110	0.022	0.170	.974	.000
M	2	452.501	226.251	1750.620	<.001	**.641**
*k*M*	2	1.663	0.831	6.430	.002	.002
*N*M*	2	20.991	10.495	81.210	<.001	.030
*k*N*M*	2	0.114	0.057	0.440	.644	.000
*D*M*	10	8.238	0.824	6.370	<.001	.012
*k*D*M*	10	2.092	0.209	1.620	.096	.003
*N*D*M*	10	1.310	0.131	1.010	.429	.002
*k*N*D*M*	10	0.299	0.030	0.230	.993	.000

Note. k = number of items, N = number of examinees, D = distribution condition, and M = model type.

Table 4.20: The Results of ANOVA for Bias of Item Difficulty Parameter Estimates

Source	*df*	Sum of Square	Mean Square	*F*	*p*	η^2
k	1	0.404	0.404	26.660	<.001	.051
N	1	0.014	0.014	0.940	.332	.002
*k*N*	1	0.029	0.029	1.940	.165	.004
D	5	0.680	0.136	8.980	<.001	.086
*k*D*	5	0.144	0.029	1.900	.093	.018
*N*D*	5	0.023	0.005	0.300	.914	.003
*k*N*D*	5	0.025	0.005	0.330	.898	.003
M	2	1.023	0.512	33.760	<.001	.130
*k*M*	2	0.054	0.027	1.780	.170	.007
*N*M*	2	0.035	0.017	1.150	.317	.004
*k*N*M*	2	0.001	0.001	0.050	.955	.000
*D*M*	10	0.074	0.008	0.540	.843	.009
*k*D*M*	10	0.016	0.002	0.130	.998	.002
*N*D*M*	10	0.014	0.002	0.100	1.000	.002
*k*N*D*M*	10	0.012	0.004	0.250	.858	.002

Note. k = number of items, N = number of examinees, D = distribution condition, and M = model type.

Recovery of Item Discrimination Parameters. Detailed results for the item discrimination parameter estimates for the Mix2PL and Mix3PL models are shown in Tables 4.21 and 4.22, respectively. Table 4.21 presents mean RMSE values for item discrimination parameter for each condition. The mean RMSE values for Mix2PL IRT model estimates of item discrimination parameter ranged from 0.108 to 0.267. Condition 1 had the highest kurtosis value and yielded the most biased item discrimination estimates for 2PL data conditions. The mean RMSE values for 2PL data conditions appeared to decrease as the number of examinees and items increased for most conditions. The Mix3PL model discrimination parameter estimates were more biased than those for the Mix2PL IRT model. Condition 3 (skewness = 1.5 and kurtosis = 2.5) yielded the most biased item discrimination estimates among the 3PL data conditions. The mean RMSE values appeared to increase as the kurtosis decreased from Condition 1 to Condition 3 and as skewness decreased from Condition 4 to Condition 6. The conditions with more items and examinees generally had lower bias values than conditions which had fewer items and examinees. Table 4.22 summarizes mean bias values of item discrimination parameter. The results of mean bias were consistent with the RMSE results for discrimination parameter estimates. The Mix3PL model estimates were more biased than those for the Mix2PL model estimates. Generally, the Mix2PL model estimates were positively biased, but the Mix3PL model estimates were negatively biased. As shown in Table 4.22, the mean bias values for item discrimination parameter estimates tended to increase as the number of examinees increased.

Table 4.21: Mean RMSE Values of Item Discrimination Parameters over 50 Replications

Condition	COND1	COND2	COND3	COND4	COND5	COND6
2PL10600	0.230	0.172	0.166	0.161	0.168	0.178
2PL28600	0.267	0.154	0.165	0.157	0.151	0.155
2PL102000	0.170	0.116	0.116	0.108	0.127	0.121
2PL282000	0.210	0.130	0.126	0.126	0.113	0.120
3PL10600	0.432	0.478	0.529	0.345	0.310	0.287
3PL28600	0.748	0.747	0.839	0.463	0.331	0.253
3PL102000	0.430	0.423	0.507	0.295	0.295	0.280
3PL282000	0.783	0.825	1.046	0.446	0.273	0.208

Table 4.22: Mean Bias Values of Item Discrimination Parameters over 50 Replications

Condition	COND1	COND2	COND3	COND4	COND5	COND6
2PL10600	−0.167	0.018	0.013	0.013	0.028	0.027
2PL28600	0.112	0.038	0.043	0.042	0.038	0.037
2PL102000	0.071	0.032	0.020	0.031	0.044	0.041
2PL282000	0.097	0.048	0.045	0.051	0.043	0.046
3PL10600	−0.157	−0.197	−0.215	−0.106	−0.053	−0.009
3PL28600	−0.163	−0.152	−0.205	−0.053	0.008	0.060
3PL102000	−0.348	−0.345	−0.394	−0.191	−0.118	−0.044
3PL282000	−0.387	−0.390	−0.504	−0.190	−0.076	0.011

ANOVA Results for Examining the Recovery of Item Discrimination Parameters. Two separate ANOVAs were performed using natural logarithm of RMSE (log[RMSE]) and square root of bias (sqrt[bias]) as dependent variables since the raw RMSE and bias for item discrimination parameter failed to meet ANOVA assumptions. The ANOVA results for log[RMSE] and sqrt[bias] of item discrimination parameter are presented in Tables 4.23 and 4.24, respectively. As shown in Table 4.23, a number of factors were significantly related to RMSE calculation for item discrimination parameter estimates ($p < .05$). However, only model type (Mix2PL vs. Mix3PL) appeared to have a large effect on the RMSE calculation. Recall that RMSE values for item discrimination parameters estimated with Mix3PL IRT model analyses were relatively larger than those for estimated item discrimination parameters in the Mix2PL model analyses. ANOVA results for log[RMSE] also provided evidence for this difference. Other significant factors for log[RMSE] include the distribution condi-

tion (D), five two-way interactions (k*N, k*D, N*D, N*M, and D*M), and one three-way interaction (N*D*M). However, all of these factors produced small to moderate effect sizes. Only the interaction of the distribution condition and model type (D*M) was found to be significant ($p < .05$) for sqrt[bias]. This interaction factor does not have much of an effect size and ($\eta^2 = .024$). As noted earlier, only factors with significant and large effect sizes were considered to be important in this study.

Table 4.23: ANOVA Results for RMSE

of Item Discrimination Parameter Estimates

Source	*df*	Sum of Square	Mean Square	*F*	*p*	η^2
k	1	0.123	0.123	1.130	.287	.000
N	1	0.027	0.027	0.250	.618	.000
*k*N*	1	0.432	0.432	3.960	.047	.001
D	5	55.933	11.187	102.770	<.001	.114
*k*D*	5	1.302	0.260	2.390	.036	.003
*N*D*	5	6.904	1.381	12.680	<.001	.014
*k*N*D*	5	0.349	0.070	0.640	.668	.001
M	1	264.735	264.735	2431.990	<.001	**.540**
*k*M*	1	0.071	0.071	0.650	.419	.000
*N*M*	1	23.693	23.693	217.650	<.001	.048
*k*N*M*	1	0.012	0.012	0.110	.743	.000
*D*M*	5	34.775	6.955	63.890	<.001	.071
*k*D*M*	5	0.428	0.086	0.790	.560	.001
*N*D*M*	5	7.180	1.436	13.190	<.001	.015
*k*N*D*M*	5	0.259	0.052	0.480	.795	.001

Note. k = number of items, N = number of examinees, D = distribution condition, and M = model type.

Table 4.24: ANOVA Results for Bias

of Item Discrimination Parameter Estimates

Source	*df*	Sum of Square	Mean Square	*F*	*p*	η^2
k	1	0.000	0.000	0.010	.918	.000
N	1	0.017	0.017	1.350	.245	.003
*k*N*	1	0.018	0.018	1.360	.244	.003
D	5	0.065	0.013	1.010	.412	.011
*k*D*	5	0.037	0.007	0.580	.717	.006
*N*D*	5	0.016	0.003	0.250	.942	.003
*k*N*D*	5	0.006	0.001	0.100	.992	.001
M	1	0.005	0.005	0.420	.516	.001
*k*M*	1	0.001	0.001	0.080	.777	.000
*N*M*	1	0.008	0.008	0.590	.443	.001
*k*N*M*	1	0.017	0.017	1.340	.247	.003
*D*M*	5	0.145	0.029	2.250	.049	.024
*k*D*M*	5	0.041	0.008	0.640	.671	.007
*N*D*M*	5	0.034	0.007	0.530	.751	.006
*k*N*D*M*	5	0.001	0.000	0.020	.979	.000

Note. k = number of items, N = number of examinees, D = distribution condition, and M = model type.

Recovery of Item Guessing Parameters. Detailed results for the item guessing estimates for the Mix3PL models are shown in Tables 4.25 and 4.26. As shown in Table 4.25, the mean RMSE values for item guessing parameter estimates were less than 0.10 for more than half of the conditions. The first three conditions yielded more biased item guessing parameter estimates. The mean RMSE tended to decrease as the skewness value decreased across the last three conditions (from Condition 4 to Condition 6). Generally, the mean RMSE values appeared to increase as the number of examinees and items increases. Similar results are presented in Table 4.26 for mean bias values. The first three conditions yield more biased estimates than the last three conditions. The mean bias values for the item guessing parameter estimates also increased as the skewness value increased across the last three conditions. However, there was no clear pattern with regard to the change in the kurtosis value across the first three conditions. The mean RMSE and bias values for item guessing parameter appeared to be lower than those for item difficulty and discrimination parameter estimates.

Table 4.25: Mean RMSE Values of Item Guessing Parameters over 50 Replications

Condition	COND1	COND2	COND3	COND4	COND5	COND6
3PL10600	0.081	0.085	0.081	0.078	0.074	0.073
3PL28600	0.109	0.108	0.119	0.089	0.089	0.073
3PL102000	0.129	0.130	0.129	0.105	0.094	0.080
3PL282000	0.160	0.164	0.188	0.115	0.085	0.066

Table 4.26: Mean Bias Values of Item Guessing Parameters over 50 Replications

Condition	COND1	COND2	COND3	COND4	COND5	COND6
3PL10600	−0.030	−0.033	−0.032	−0.027	−0.023	−0.021
3PL28600	−0.049	−0.049	−0.055	−0.037	−0.031	−0.024
3PL102000	−0.060	−0.061	−0.060	−0.046	−0.039	−0.028
3PL282000	−0.078	−0.080	−0.092	−0.054	−0.038	−0.026

ANOVA Results for Examining the Recovery of Item Guessing Parameters. Two separate ANOVAs were performed using natural logarithm of RMSE (log[RMSE]) and raw data of bias as dependent variables. ANOVAs for item guessing parameter do not include a factor called model type as estimates from only Mix3PL model include this parameter. The ANOVA tables for log[RMSE] and bias of item guessing are presented in Tables 4.27 and 4.28, respectively. As shown in Table 4.27, a number of factors were significant, including the number of items (k), sample size (N), distribution condition (D), and two two-way interactions (k*N and k*D). Only the number of items, however, had a large effect on the RMSE calculation for the item guessing parameter. Only distribution condition and sample size were significant, albeit with small effect sizes, for bias of the item guessing parameter. There appears to be meaningful differences in the RMSE among the two test-length conditions as is evidenced by a large and significant effect size ($\eta^2 = .255$).

Table 4.27: The Results of ANOVA for RMSE

of Item Guessing Parameter Estimates

Source	*df*	Sum of Square	Mean Square	*F*	*p*	η^2
k	1	55.840	55.840	164.990	<.001	**.255**
N	1	6.388	6.388	18.870	<.001	.029
*k*N*	1	4.133	4.133	12.210	.001	.019
D	4	17.799	3.560	10.520	<.001	.081
*k*D*	4	11.107	2.221	6.560	<.001	.051
*N*D*	4	3.515	0.703	2.080	.068	.016
*k*N*D*	4	2.207	0.441	1.300	.262	.010

Note. k = number of items, N = number of examinees, and D = distribution condition.

Table 4.28: The Results of ANOVA for Bias of Item Guessing Parameter Estimates

Source	*df*	Sum of Square	Mean Square	*F*	*p*	η^2
D	5	0.059	0.012	2.350	.040	.025
k	1	0.014	0.014	2.740	.099	.006
N	1	0.038	0.038	7.620	.006	.016
*D*k*	5	0.007	0.001	0.290	.916	.003
*D*N*	5	0.010	0.002	0.400	.849	.004
*k*N*	1	0.000	0.000	0.000	.963	.000
*D*k*N*	5	0.001	0.000	0.030	1.000	.000

Note. k = number of items, N = number of examinees, and D = distribution condition.

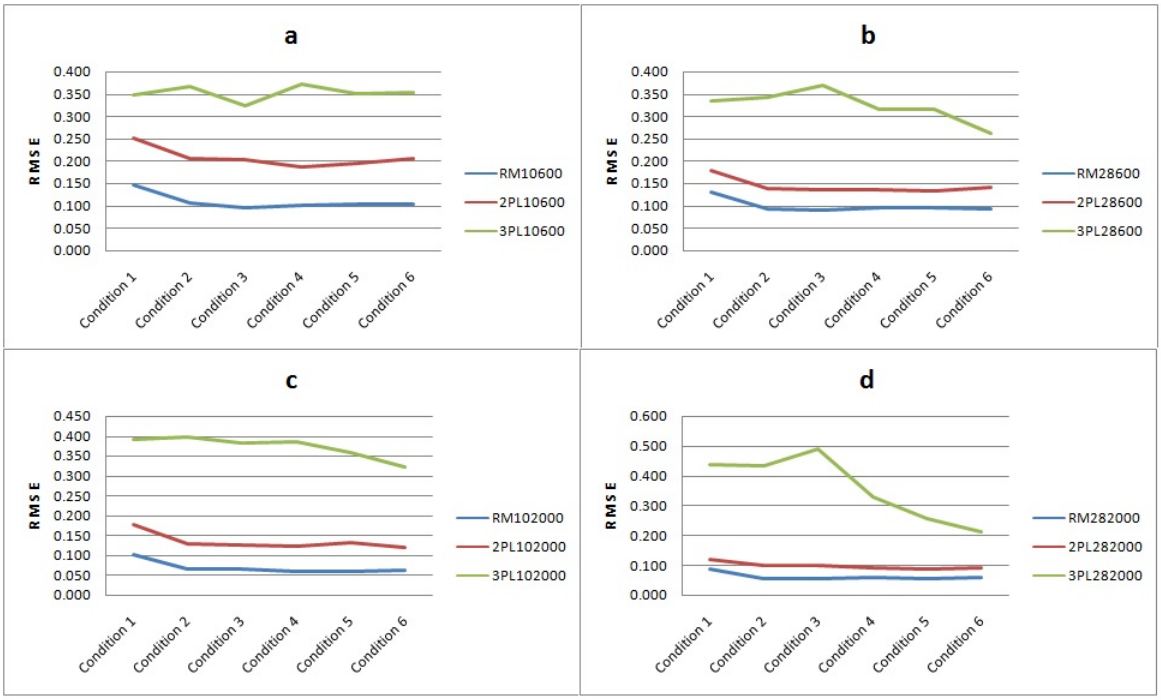

Figure 4.6: RMSE Plot for Item Difficulty Parameter Estimates

Summary of the Recovery Assessment of Simulation Study 2. Two evaluative criteria were used in this simulation study for the item parameter estimates: the RMSE and bias of the estimates. The RMSE results for item difficulty parameter estimates are plotted for each data condition in Figure 4.6. The results of Simulation Study 2 showed similar patterns to that of the Simulation Study 1. For example, the difficulty parameter estimates from the MRMs tended to produce smaller RMSE and bias values than item difficulty estimates from the Mix2PL and Mix3PL IRT model analyses (see Figure 4.6). As shown in Figure 4.6, the accuracy of the item difficulty parameter appeared to decrease as the model complexity increased. Apart from the model type, the distribution condition also appeared to have an important effect on the recovery of item difficulty parameter. It was more difficult to recover item difficulty parameters in Condition 1 than in the other

conditions (see Figure 4.6). The increase in the kurtosis values appeared to increase the RMSE and bias values for the item difficulty parameter estimates while skewness did not show any clear pattern. ANOVA analyses also showed that the model type and distribution conditions were important for RMSE and bias calculation for item difficulty parameter.

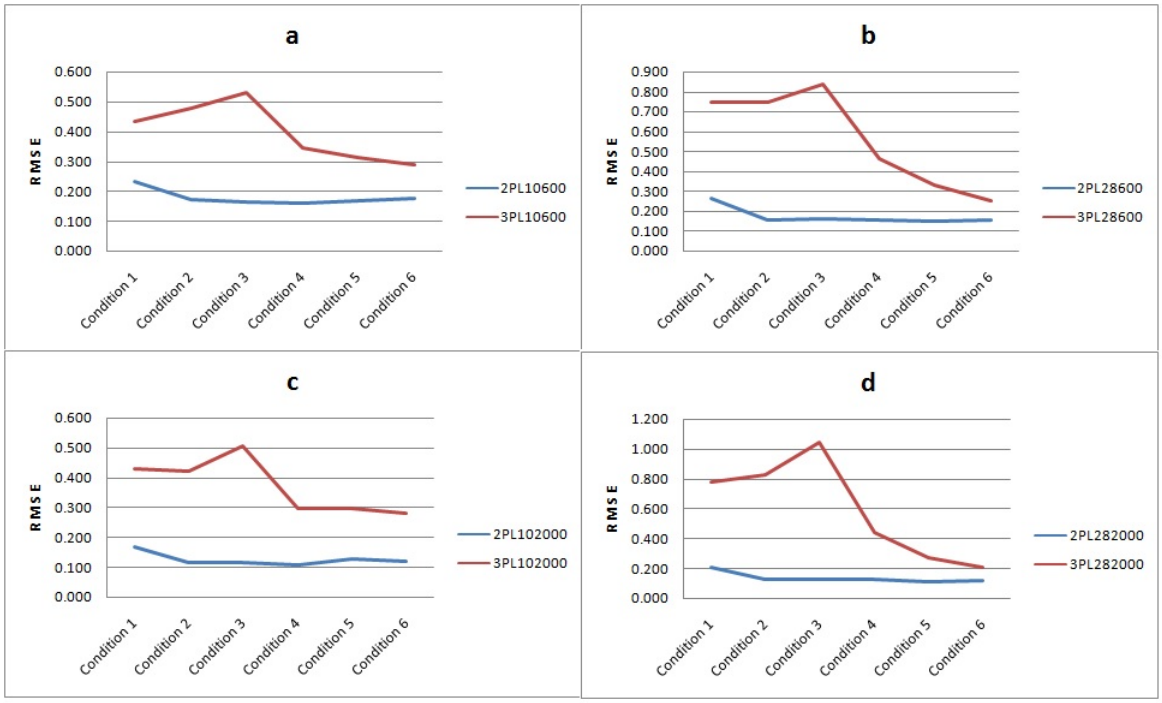

Figure 4.7: RMSE Plot for Item Discrimination Parameter Estimates

The RMSE results for item discrimination parameter estimates are plotted in Figure 4.7. For the discrimination parameter, conditions with the highest kurtosis values yielded the highest RMSE and bias values. Mix3PL analyses in Condition 3 yielded the most biased estimates (see Figure 4.7). The model type appeared to be an important effect for RMSE and bias calculation for item discrimination parameters across the two MixIRT models. The Mix2PL model analyses yielded fewer biased item discrimination estimates than Mix3PL model analyses. Consistent with the results for item difficulty, ANOVA results of RMSE

values showed that the model type and distribution conditions were the most important factors for RMSE calculation of the item discrimination parameter.

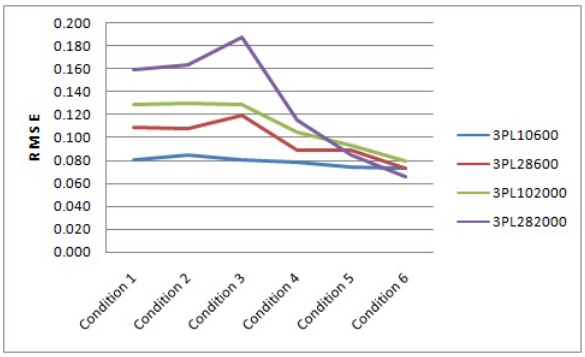

Figure 4.8: RMSE Plot for Item Guessing Parameter Estimates

The RMSE results for item guessing parameter estimates are plotted for Mix3pl IRT data conditions in Figure 4.8. As with the first simulation study, almost all of the conditions yielded very small RMSE (<0.10) and bias (<0.02) values for item guessing parameter estimates. Mix3PL model analyses had difficulty producing accurate item difficulty and discrimination estimates. Results were better for estimating the guessing parameter. As shown in Figure 4.8, the last two data conditions (Condition 5 and Condition 6) yielded the least biased guessing parameter estimates. ANOVA analyses with RMSE values showed that the number of items had an important effect on the recovery of the item guessing parameter in the second simulation study.

4.3.2 Correct Positive Rates

As with the first simulation study, the proportion of correct positives for three MixIRT models were calculated based on minimum AIC and BIC between one-class and two-class solutions in this simulation study. The number of classes for the given data set was defined as correct when the considered information index for a one-class solution was smaller than that of the two-class solution. These proportions are presented in Tables 4.29, 4.30, and 4.31

for each condition of the MRM, Mix2PL and Mix3PL IRT models, respectively. As stated above, condition names that are given in the first column of Tables 4.29 to 4.31 include model name, number of items and number of examinees. For example, the condition 2PL282000 indicates a data condition that was generated with the 2PL IRT model for 28 items and 2,000 examinees. Also note that six different distribution conditions were generated in this simulation study. The first three simulation conditions were generated to have the same skewness value (i.e., 1.5) and different kurtosis values. Kurtosis values were 4.0, 3.5, and 2.5 for Condition 1, Condition 2, and Condition 3, respectively. The last three conditions were generated to have the same kurtosis values (i.e., 3.5) and different skewness values. The skewness values were 1.0, 0.5 and 0 for Condition 4, Condition 5, and Condition 6, respectively.

Table 4.29: The Correct Positive Rates for MRM Analyses over 50 Replications

Condition	COND1		COND2		COND3		COND4		COND5		COND6	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
RM10600	6	42	1	40	2	36	4	42	6	41	9	47
RM102000	0	15	0	3	0	2	0	31	0	35	2	35
RM28600	40	50	20	50	14	50	41	50	43	50	45	49
RM282000	0	50	0	50	0	45	2	50	12	50	25	50

The correct positive rates for MRM analyses are presented in Table 4.29. BIC correctly detected the correct model in all replications for the MRM analyses with 28-item conditions except for Condition 3 and Condition 6. The correct positive rates for the BIC index were low under the 10-item and 2,000-examinee conditions. The performance of the BIC index

appeared to increase as kurtosis increased across the first three distribution conditions from Condition 3 to Condition 1. The performance of the BIC index appeared to be lower under the first three conditions than the performances under the last three conditions. The correct positive rates for the BIC were lowest for Condition 2 (3 out of 50 correct model selections) and Condition 3 (2 out of 50 correct model selections) for 10-item and 2,000-examinee conditions. Overall, the performance of the AIC index was lower than the BIC index for the MRM analyses. According to the results from the AIC index, detection of spurious latent classes appeared to be a problem for conditions with the extreme skewness and kurtosis values generated in this study. Correct positive rates for the AIC index were low in the 10-item conditions for the MRM. AIC index was generally higher only in the 28-item and 600-examinee conditions. The correct positive rates for the AIC index appeared to increase as the number of examinees decreased, and the sample size increased.

Table 4.30: The Correct Positive Rates for Mix2PL Analyses over 50 Replications

Condition	COND1		COND2		COND3		COND4		COND5		COND6	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
2PL10600	3	30	2	29	0	20	9	40	16	45	12	45
2PL102000	0	2	0	1	0	0	0	17	6	42	12	44
2PL28600	19	50	10	50	1	50	42	50	47	50	47	50
2PL282000	0	29	0	33	0	26	6	48	29	49	45	50

Table 4.30 presents the correct positive rates for Mix2PL model analyses for the six distribution conditions. For all of the 28-item and 600-examinee conditions, BIC correctly identified the generating model in all 50 replications. The correct positive rates for BIC index

were lower, however, for the Mix2PL model under the first three distribution conditions. The BIC detection rate was higher, however, the remaining three conditions. The correct positive rate for BIC appeared to increase as the kurtosis value decreased across the first three conditions (Condition 1 to Condition 3). No clear pattern existed with regard to increase or decrease of skewness across the last three conditions in terms of correct selection of the BIC. AIC was less effective at selection of the correct model except for the 28-item and 600-examinee condition. The correct positive rate of the AIC was lower for the first three conditions. Apparently, increase in the kurtosis values lead to detection of spurious latent classes. The correct positive rates for AIC for Mix2PL analyses appeared to increase as the number of items increased and as the sample size decreased. As with the results of the MRM analyses, overall performance of the AIC index was lower than that of the BIC index for Mix2PL analyses. Based on the results of the AIC index, latent nonnormality appeared to yield to detection of spurious latent classes. The results based on the BIC index showed fewer spurious latent class problems for Mix2PL analyses. Only the first three data conditions (Condition 1 to Condition 3) with 10 items and 2,000 examinees resulted in the selection of two-class solutions based on the BIC index.

Table 4.31: The Correct Positive Rates for Mix3PL Analyses over 50 Replications

Condition	COND1		COND2		COND3		COND4		COND5		COND6	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
3PL10600	21	50	31	50	33	50	15	48	15	50	12	49
3PL102000	3	48	5	47	5	46	1	45	4	46	1	37
3PL28600	46	50	49	50	50	50	50	50	49	50	50	50
3PL282000	32	50	34	50	22	50	46	50	48	50	43	50

Table 4.31 presents the correct positive rates for Mix3PL IRT model analyses. In almost all distribution conditions, BIC supported selection of one class. The BIC index selected the correct model for more than 90% of the replications in all of the distribution conditions except Condition 6, with skewness of 0 and kurtosis of 3.5 (i.e., Condition 6). Condition 6 with 10 items and 2,000 examinees yielded the lowest result (37 out of 50) in terms of the BIC index. The conditions with 28 items showed 100% correct detection with the BIC index. For the 10-item conditions, the correct positive rate of the BIC decreased as the number of examinees increased. The number of correct selections was higher for AIC for the Mix3PL model compared to the previous models. Consistent with previous results, however, the number of correct selections by AIC was lower than for BIC. Further, AIC had lower numbers of correct detections in all of the distribution conditions with 10 items and 2,000 examinees. Performances of the AIC index appeared to improve as the number of items increased and as the number of examinees decreased. Consistent with the results of MRM and Mix2PL IRT model analyses, overall performance of the AIC index was lower than the performance of the BIC index. The first three distribution appeared to yield more spurious latent class problems than last three distribution conditions. Apparently, the Mix3PL models were more robust to latent non-normality than either the MRM or Mix2PL models based on results for both the AIC and BIC.

Summary of the Correct Positive Rates of Simulation Study 2. For the second simulation study, the first three distribution conditions yielded more spurious latent class problems than the last three conditions for the three MixIRT model analyses. Therefore, the results of the first three distribution conditions did not select the correct solution for the most of the sample size and number of items conditions in terms of the AIC index. Overall performance of the AIC index was lower than that of the BIC index. The performances of both indices appeared to increase as the number of items increased. However, the correct positive rates decreased as the number of examinees increased. A pattern appeared to exist

with regard to model complexity. Mix2PL models appeared to be more robust to latent non-normality than MRMs, in turn, Mix3PL models were more robust than Mix2PL models.

4.4 Results of Empirical Data

In this section, the effect of a non-normal ability distribution is illustrated on a sample data set from a large testing program, the TIMSS (Foy, Arora, & Stanco, 2013). The first part in this section describes the data that were used for the MixIRT analysis. Next, some preliminary analyses are presented in terms of number of latent class, unidimensionality, and distributional characteristics. The last part of this section presents the results of MixIRT analyses using MCMC algorithm.

4.4.1 Data

Data were taken from the TIMSS 2011 8th grade mathematics test. The TIMSS 2011 assessment included 14 student achievement booklets and 28 blocks (14 mathematics and 14 science). Each student took only one booklet which included two mathematics and two science tests. Each booklet was composed of similar proportions of item types, including multiple choice, short answer, and extended response items. In addition, each booklet was arranged to have similar proportions of content in the domains of number, algebra, geometry, and data and the cognitive domains of knowing, applying, and reasoning. Six blocks were made publicly available for TIMSS 2011 assessment. Only four booklets (Booklet 1, Booklet 2, Booklet 5, and Booklet 6) can be obtained using pairs of the available blocks. Administration of Booklet 5 to Korean students was selected for the MixIRT analyses.

Booklet 5 was composed of Block 5 and Block 6 with 32 items including multiple choice, short answer, and extended response items. Of those, 18 multiple-choice items were used for the example in this study. There were 369 examinees who had taken Booklet 5. The Korean sample was the highest performing sample and was included in this study to provide

a data set with a negatively skewed ability distribution. The original TIMSS data set was analyzed with a 3PL IRT model and item parameter estimates were reported for each item at the TIMSS web site. The mean test score for all Korean students was 613, and the overall average of the participating countries was 500.

4.4.2 Preliminary Analyses

Before proceeding to MixIRT analyses, it was necessary to investigate the data to determine whether it was unidimensional, had a single class and had a non-normal ability distribution. First, the dimensionality of the data set was assessed using TESTFACT software package. TESTFACT was selected because our data set was composed of only dichotomous items and this software was also used in simulation studies in this study. As mentioned previously, the chi-square difference test and the proportion of variance accounted for approaches are available from the TESTFACT software. In order to calculate chi-square difference test statistic and eigenvalues, two EFAs were employed by setting the number of factors to one and two. Guessing parameter was also set to .25 in the TESTFACT syntax. The chi-square values and degrees of freedom (df) values for each analysis were used to calculate the chi-square and df values for the chi-square difference test. Using the calculated chi-square test statistic, the null hypothesis and the alternative hypothesis are presented below:

H_0 : 1-factor model provides an adequate fit to the data.

H_a : 2-factor model provides an adequate fit to the data.

The chi-square difference test statistic yielded a value of 33.28 with df=17 (p=.003). This result suggested the alternative hypothesis of non-normality was more supported. As noted earlier, the chi-square difference test may not be reliable in situations in which small numbers of empirical frequencies appear in some response patterns. In the case of lack of support from chi-square difference test, the eigenvalues can be used to investigate the proportion of variance accounted for the first factor. The proportion of variance accounted for

approach showed that the first factor explained 41.7% (i.e., 7.52 out of 18) of the variation in the data set (see Figure 4.9). Based on Reckase's (1979) 20% criterion, this data set was determined have a dominant first factor structure and, therefore, to be unidimensional.

```
PHASE 6:  FACTOR ANALYSIS
KOREAN TIMSS DATA
NON-ADAPTIVE FULL-INFORMATION ITEM FACTOR ANALYSIS
------------------------------------------------------------
DISPLAY   2.   THE POSITIVE LATENT ROOTS OF THE CORRELATION MATRIX
             1         2         3         4         5         6
         7.521330  1.535346  1.260061  1.135829  1.002908  0.893594
             7         8         9        10        11        12
         0.782123  0.698084  0.671704  0.616699  0.517880  0.417175
            13        14        15        16        17
         0.331778  0.309256  0.177240  0.157874  0.041616
```

Figure 4.9: Positive Latent Roots of the Correlation Matrix Output From TESTFACT

Second, the WINMIRA software package was used to examine whether the Korea TIMSS data came from a single-class. WINMIRA software was selected because it allows us to estimate the MRM with conditional maximum likelihood (CML) estimation. CML estimation was used for this analysis as it provides results which are independent from distributional characteristics of the ability parameter. An iterative EM algorithm was employed to identify latent classes, and the CML estimation was used to obtain item and person parameters in WINMIRA. Several MRM models were fit to the data, each were conducted with varying number of latent classes from one to five, because the latent classes were not known beforehand. The best fitting model was identified based on the smallest indices that were provided

by WINMIRA. The summary of two fit indices (AIC and BIC) and the loglikelihood (i.e., LL) values for one- to five-class solutions are presented in Table 4.32. As shown in Table 4.32, AIC selected the four-class solution whereas the BIC index selected the one-class solution. As is the case with simulation study and previous research, AIC index selected the model with the larger number of latent classes. Based on the BIC result, the Korean sample data set was determined to have a single-class solution.

Table 4.32: Model Fit Statistics for Winmira Analyses of Korean TIMSS Data

Model	LL	AIC	BIC
1-Class	−2,841.07	5,720.15	**5,794.45**
2-Class	−2,788.30	5,654.59	5,807.12
3-Class	−2,761.50	5,640.99	5,871.73
4-Class	−2,735.71	**5,629.42**	5,938.37
5-Class	−2,723.03	5,644.05	6,031.22

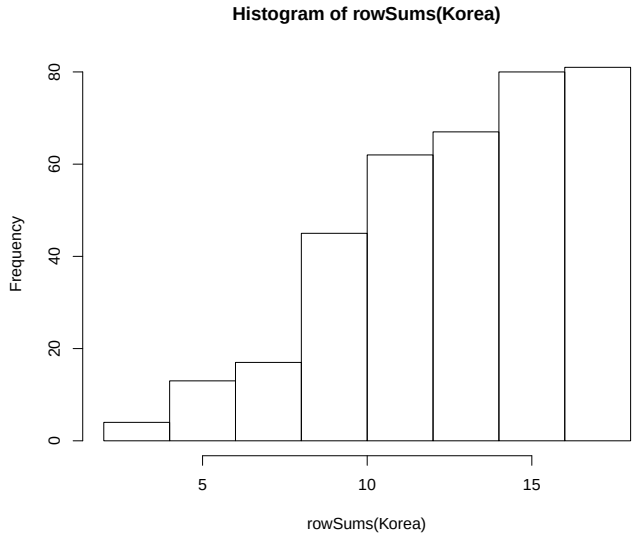

Figure 4.10: Histogram of Sum of Raw Scores

For Korean Sample Data (N=369, k=18)

Third, the distribution of the ability parameter of the Korean sample data was examined to determine whether or not it had a non-normal ability distribution. Before proceeding to characteristics of latent ability distribution, we first describe what the empirical distribution of the data set looked like. The frequency of the total raw scores is presented in Table 4.33. These scores are plotted in Figure 4.10. As was expected, the total raw scores for the most of the students were distributed on the right side of the plot (see Figure 4.10). Additionally, more than 80% percent of the students scored higher than the mean score of 8 (see Table 4.33). Thus, this data produced a negatively-skewed empirical distribution.

Table 4.33: Frequency of the Raw Scores for Korean Students with Booklet5

Raw Score	Frequency	Cumulative Frequency
3	1	1
4	3	4
5	5	9
6	8	17
7	5	22
8	12	34
9	26	60
10	19	79
11	29	108
12	33	141
13	36	177
14	31	208
15	37	245
16	43	288
17	46	334
18	35	369

The empirical distribution appears to have a non-normal, skewed distribution. This does not guarantee that the latent ability distribution is also distributed non-normally. Fortunately, it is possible to detect non-normality of the latent ability distribution with a method called Ramsay-curve IRT (RC-IRT) analysis. The RC-IRT method estimates the latent den-

sity as a B-spline-based density (Woods, 2008; Woods & Thissen, 2006) simultaneously with estimation of item parameters. This method can be employed with a software package called RCLOG version 2 (Woods, 2006b).

RCLOG program was designed to detect and correct non-normality in the data set. This software was shown to perform well for estimation of IRT models for data sets with non-normal latent ability distributions (Woods, 2006; Woods & Thissen, 2006). Item and ability parameters are estimated through a combination of Bock-Aitkin IRT with Ramsay curves (Woods & Thissen, 2006). The RCLOG program produces a series of Ramsay curves for the latent distribution based on different number of breaks (two to six) and orders (two to six). Breaks refer to the locations on the horizontal axis where different polynomials are joined together while order denotes the order of the polynomial B-splines (Woods, 2006, p. 255). As the number of breaks and order increased the number of parameters and the flexibility of the density increase (Woods, 2006). As suggested in RCLOG manual, the numbers from two to six are used for breaks and order in this study. That is, we obtain 25 combinations of breaks and order (e.g., 2-2, 2-3, 6-6). The best model is selected from among the candidate models (i.e., 25 Ramsay curves) considered using the fit indices that are provided by the RCLOG software. The available fit indices on the RCLOG software are AIC, BIC, and the Hannan-Quinn (HQ; Hannan, 1987) criterion. A more detailed explanation of RCLOG software and RC-IRT method can be found in Woods and Thissen (2004), and Woods (2006b).

The 3PL model was fit to responses to the 18 items on the Korea TIMSS data set. All possible RC-IRT models with order up to five and number of breaks up to five were fit using the RCLOG software package (see Figure 4.11). As shown in the examples in the RCLOG manual, 121 quadrature points were specified ranging from -6 to 6. The SD on the prior for the spline coefficients was 75. Graphical representations of possible latent ability distributions are plotted in Figure 4.11. Results of fit statistics for each break-order pair and corresponding skewness and kurtosis values are presented in Table 4.34. The Ramsay curve

with three breaks and order of two has the smallest AIC, BIC, and HQ values. Based on the minimum fit index criteria, the Ramsay curve with three breaks and order of two was selected as the best model. This model indicated that the latent ability distribution had a skewness value of −1.21 and a kurtosis value of 2.53 (see bolded row in Table 4.34).

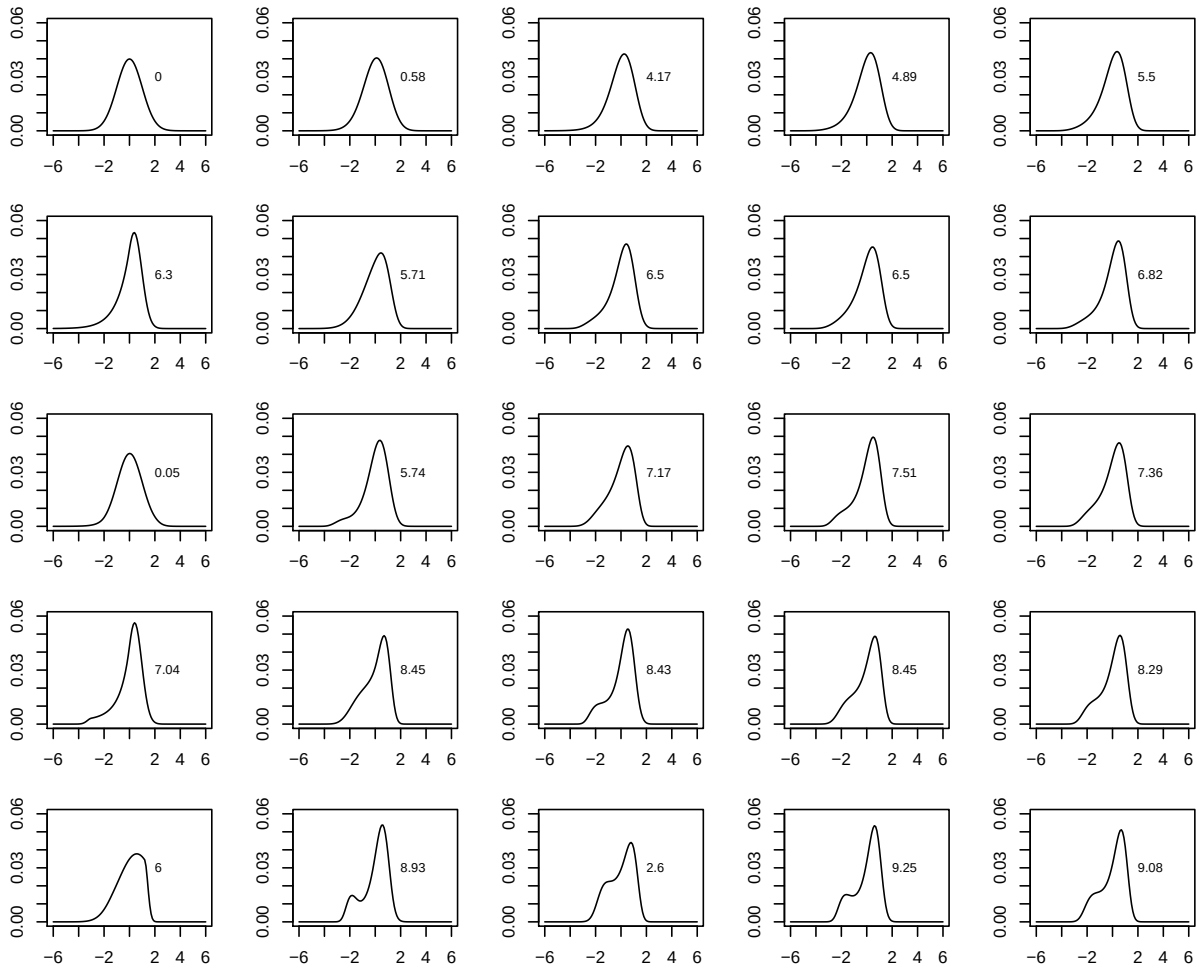

Figure 4.11: Candidate Ramsay Curves for TIMSS Items (N=369, k=18)

Note. Elements of each column from top to bottom represent break points from two to six, and elements of each row from left to right represent order numbers from two to six. Numbers in each box correspond to $-2(likelihood_{2-2} - likelihood_{other})$

Table 4.34: Fit Indices From 25 Ramsay Curves

Breaks	Order	AIC	BIC	HQ	Skewness	Kurtosis
2	2	2347.29	2355.12	2350.40	0.00	0.00
2	3	2348.72	2360.46	2353.38	−0.23	0.17
2	4	2347.13	2362.78	2353.34	−0.62	0.87
2	5	2348.41	2367.97	2356.18	−0.66	0.67
2	6	2349.79	2373.28	2359.12	−0.68	0.55
3	**2**	**2342.99**	**2354.73**	**2347.66**	**−1.21**	**2.53**
3	3	2345.59	2361.24	2351.80	−0.61	0.30
3	4	2346.80	2366.36	2354.57	−0.78	0.51
3	5	2348.79	2372.27	2358.12	−0.73	0.39
3	6	2350.48	2377.87	2361.36	−0.91	0.74
4	2	2351.24	2366.89	2357.46	−0.17	0.35
4	3	2347.56	2367.12	2355.33	−0.89	0.95
4	4	2348.12	2371.60	2357.45	−0.66	−0.01
4	5	2349.78	2377.18	2360.67	−0.83	0.30
4	6	2351.93	2383.24	2364.37	−0.70	0.06
5	2	2346.25	2365.82	2354.02	−1.10	1.22
5	3	2346.84	2370.32	2356.17	−0.68	−0.22
5	4	2348.87	2376.26	2359.75	−0.81	0.02
5	5	2350.85	2382.16	2363.28	−0.69	−0.22
5	6	2353.01	2388.23	2367.00	−0.72	−0.13
6	2	2349.29	2372.78	2358.62	−0.58	−0.13
6	3	2348.37	2375.76	2359.25	−0.77	−0.23
6	4	2356.70	2388.00	2369.13	−0.42	−0.77
6	5	2352.05	2387.27	2366.04	−0.71	−0.37
6	6	2354.22	2393.35	2369.76	−0.67	−0.41

4.4.3 MixIRT Results

In this part, we illustrate MCMC estimation of three MixIRT models by fitting them to the Korean sample data set. The MCMC algorithm was implemented in WinBUGS as described in simulation studies. The following priors and hyper-priors were used for the MRM:

$$\begin{aligned}
\beta_{ig} &\sim \text{Normal}(0,\ 1),\ i = 1, ..., I, g = 1, ..., G, \\
\theta_j &\sim \text{Normal}(\mu(\theta)_g, 1),\ j = 1, ..., J, \\
\mu(\theta)_g &\sim \text{Normal}(0,\ 1),\ g = 2, ..., G, \mu_1 = 1, \\
g &\sim \text{Multinomial}(1, \pi_g[1 : G])
\end{aligned}$$

where J is the total number of examinees, I is the total number of items, and G is the number of latent classes. θ_j is the ability parameter for person j, and β_{ig} is the difficulty parameter of item i within class g. As was done in Bolt et al. (2002), the standard deviation for each latent class, σ_g, was fixed at 1 for each group. A Dirichlet distribution with equal proportions was used as the prior for the group membership parameter. For example, a Dirichlet distribution with (.5, .5) was used as the prior for π_g for the two-class model and (.5, .5, .5) was used as the prior for the three-class model. In addition, .5 was used as initial values for the mixture proportions. Two additional priors were used in the Mix2PL and Mix3PL model analyses:

$$\begin{aligned}
\alpha_{ig} &\sim \text{Normal}(0,\ 1)I(0,) \\
\gamma_{ig} &\sim \text{Beta}(5,\ 17).
\end{aligned}$$

As was done in Simulation Studies 1 and 2, the numbers of burn-in and post-burn-in iterations were 6,000, 7,000, and 9,000 for the MRM, Mix2PL, and Mix3PL, respectively. Each of the three MixIRT models were estimated with from one to five latent classes. The best fitting solution was identified based on information criteria indices, AIC and BIC. Results for fit indices of the three MixIRT models are presented in Table 4.35. As can be seen in this table, AIC selected the three-class solution and BIC selected the two-class

solution for the MRM. Both AIC and BIC selected the one-class solution for the Mix2PL and Mix3PL models. As was the case with simulation studies and previous research, AIC may be overestimating the number of latent classes. Thus, the two-class solution was identified as the best fit for the MRM and the one-class solution as the best fit for the Mix2PL and Mix3PL models based on the BIC index. These results appear to be consistent with results from the two simulation studies in this dissertation with respect to performances of the AIC and BIC indices.

Table 4.35: MCMC-Based Fit Statistics for MixIRT Analyses of Korea TIMSS Data

Model	MRM		Mix2PL		Mix3PL	
	AIC	BIC	AIC	BIC	AIC	BIC
1-Class	5,237.00	5,312.00	**5,074.00**	**5,219.00**	**5,058.00**	**5,273.00**
2-Class	5,027.00	**5,179.00**	5,085.00	5,378.00	5,155.00	5,590.00
3-Class	**5,015.00**	5,246.00	5,075.00	5,516.00	5,248.00	5,901.00
4-Class	5,046.00	5,355.00	5,220.00	5,810.00	5,369.00	6,242.00
5-Class	5,082.00	5,469.00	5,292.00	6,031.00	5,415.00	6,506.00

Estimates of item parameters for the selected models are presented in Table 4.36. For the MRM, one unique set of item difficulty (b) parameter estimates is provided for each of two classes (i.e., b_1 and b_2). For the one-class Mix2PL model, one set of estimates is provided for item difficulty (b) and discrimination (a). For the Mix3PL, item difficulty, item discrimination, and item guessing (c) parameter estimates are reported for the one-class model. Mixture proportion estimates in two-class MRM estimation were .51 and .49, respectively for Class 1 and Class 2.

Table 4.36: Item Parameter Estimates from the MixIRT Analyses of the Korean Sample

	MRM		Mix2PL		Mix3PL		
Items	b_1	b_2	a	b	a	b	c
Item 1	−1.331	−1.515	1.311	−2.253	1.300	−1.969	0.276
Item 2	2.310	1.645	1.909	0.251	2.386	0.426	0.096
Item 3	0.656	0.155	1.686	−0.688	1.955	−0.386	0.187
Item 4	1.360	−0.406	2.126	−0.426	2.816	−0.180	0.155
Item 5	0.483	0.597	1.372	−0.764	1.439	−0.513	0.161
Item 6	0.141	1.067	0.836	−1.104	1.120	−0.319	0.279
Item 7	−0.593	−0.855	1.555	−1.526	1.637	−1.247	0.231
Item 8	2.333	2.225	1.249	0.564	2.501	0.781	0.141
Item 9	−0.866	−0.236	1.250	−1.767	1.213	−1.480	0.253
Item 10	−1.151	0.238	0.959	−2.153	0.953	−1.711	0.271
Item 11	−1.849	−1.862	1.850	−2.203	1.838	−2.102	0.232
Item 12	−1.065	−2.047	2.347	−1.594	2.353	−1.462	0.233
Item 13	0.372	0.294	1.264	−0.921	1.653	−0.403	0.252
Item 14	−0.554	−1.774	2.046	−1.400	2.274	−1.161	0.228
Item 15	−0.991	−0.902	1.480	−1.803	1.458	−1.607	0.226
Item 16	0.990	−0.142	1.821	−0.561	2.403	−0.245	0.188
Item 17	1.479	2.998	0.595	0.971	1.276	1.434	0.201
Item 18	−1.984	−1.812	1.882	−2.255	1.889	−2.129	0.252

Chapter 5

DISCUSSION

5.1 Summary

This study investigated the effect of non-normal latent ability on the number of latent classes extracted in MixIRT models. Two simulation studies were conducted to determine the effects of different types of distributions and different amounts of kurtosis and skewness for different length tests and different sample sizes on detection of latent classes. In addition, an empirical study was presented illustrating how the effects of latent nonnormality might be detected in real data.

Five different distribution conditions, two test lengths, and two sample sizes were manipulated in Simulation Study 1. Distribution conditions included normal, platykurtic, skewed, uniform and bimodal symmetric. Simulation Study 2 focused on examining the effects of six extreme non-normality conditions that might be realized in practical testing conditions. These six conditions were simulated by manipulating different combinations of skewness and kurtosis values along with two test lengths, two sample sizes, and three different MixIRT models. All of the data sets in Simulation Studies 1 an 2 were generated to have a single class and were analyzed with one- and two-class solutions. The best solution was determined based on AIC and BIC values for each MixIRT model. Finally, data from a large-scale testing program were used to demonstrate how to detect non-normality and how to estimate the effects of non-normality on the number of latent classes.

For the MRM analyses conducted in Simulation Study 1, the two-class MixIRT model was consistently found to be a better representation of the data than the one-class model under

both bimodal and uniform data conditions. As expected, the MRM analyses of the data with normal and both of the typical non-normal ability distributions, i.e., the skewed and platykurtic distributions, showed fewer over-extraction problems based on the BIC index. The correct positive selection rates for the normal and the two typical non-normal distributions were very low based on the AIC index. Thus, the overall performance of AIC appeared to be worse than that of the BIC in the context of latent non-normality. This appears to be consistent with results from previous research suggesting that AIC tends to select the more complex model, in this case, the model with more latent classes (e.g., Li et al., 2009).

The results for the Mix2PL and Mix3PL model analyses showed similar patterns in terms of relative performance of the AIC and BIC information criterion indices. AIC selected models with two-classes more than did BIC. As was the case for the MRM, this result was consistent with previous research on model selection that found AIC to select more complex model solutions. For most of the simulated conditions, unlike results for the MRM, non-normality did not appear to lead to over-extraction for either the Mix2PL or MiX3PL models. More complex models such as the Mix2PL and Mix3PL, in other words, appear to be more robust to violation of latent normality in that both tended to yield fewer spurious latent class solutions.

The recovery of the generated parameters was assessed in each simulation study by calculating RMSE and bias statistics for each condition. In Simulation Study 1, recovery of parameter estimates from the MRMs was generally better than that for the Mix2PL model estimates and recovery estimates for the Mix2PL model were better than for the Mix3PL model. Recovery of generated parameters also varied across different distribution conditions. In Simulation Study 1, recovery was less accurate in the bimodal symmetric and uniform distribution conditions than for the normal, platykurtic and skewed distributions.

The results for Simulation Study 2 showed similar patterns to those of Simulation Study 1. For example, overall performance of the AIC index was poorer than that of BIC. Recovery

declined as model complexity increased as shown by increases in RMSE and bias as the model complexity increased. In addition, the more complex Mix3PL model yielded fewer spurious latent class detections than the simpler MRM. Results suggested that latent non-normality may be causing extraction of spurious latent classes and to some extent the Mix2PL model. Results based on both AIC and BIC suggest that over-extraction is likely for the MRM.

The results of the empirical study also showed some similarities to the results from the two simulation studies. The Korean sample data set from TIMSS mathematics assessment used for the real data example appeared to have a non-normal, skewed ability distribution. This skewed data set was analyzed with three MixIRT models with solutions for from one to five classes. Two fit indices (i.e., AIC and BIC) were used to identify the best solution for the three MixIRT analyses. For the MRM analyses, the two-class model appeared to be best fitting solution despite the fact that only one class was present based on analyses with the CML, a distribution-free estimation method. The Mix2PL and Mix3PL indicated only a single class model based on both AIC and BIC indices. As with the two simulation studies, more complex models appeared to be more robust to violations of latent non-normality. Skewness caused extraction of a spurious latent class only for the MRM.

5.2 Conclusion

A few important conclusions can be made based on the findings. First, results suggested that latent non-normality may be capable of causing extraction of spurious latent classes for the MRM analyses. More complex models, however, such as the Mix2PL and Mix3PL appeared to be more robust to latent non-normality in that both tended to yield fewer spurious latent class solutions. The robustness of the Mix2PL and Mix3PL models to latent non-normality could be due to effect of penalization used to calculate the AIC and BIC. With respect to the penalty term used in the information indices considered here, the more parameters added to the model, the larger the penalty term.

Second, the performance of fit indices in this study was consistent with previous research on model selections (e.g., Li et al., 2009; Preinerstorfer & Formann, 2011), which have shown that AIC tended to select the more complicated MixIRT model and BIC the correct MixIRT model compared to AIC. The performance of these information indices used to determine model fit also may be a function of the underlying distribution of the data. Although the results of the two simulation studies demonstrated spurious latent class problems, it was not clear whether the underlying non-normality caused over-extraction or whether the fit indices overestimated the correct model solutions.

Third, interpretability of the latent classes in any model needs to be a consideration in determining model selection. Relying only on statistical criteria may not always yield interpretable solutions. Results in this study suggested that it may be misleading, even under the most ideal conditions, to use the AIC index for identifying number of latent classes. Thus, the solution accepted should be expected to have sufficient support not only from statistical criteria but also from the interpretability of the classes (Bauer & Curran, 2003).

Finally, results of this study suggest that latent non-normality not only caused over-extraction but also resulted in poor estimation of the item parameters. When dealing with empirical data, as opposed to simulated data, it is advisable to test the distribution of ability for the possibility of non-normality. The RC-IRT method provides a useful tool for this kind of analysis. Final model selection also should be interpreted cautiously when severe non-normality is present. Several methods have been proposed for the correcting the non-normality in the traditional IRT analyses (Bock & Aitkin, 1981; Mislevy & Bock, 1990; Woods, 2004). No solution has been proposed, however, for handling this problem in the context of MixIRT model estimation. It is also important that model selection should include interpretability of the results rather than solely on statistical evidence.

5.3 Future Research

As with any simulation study, generalization based on the results of this dissertation is limited to the number of conditions manipulated here. The conditions simulated here were based on practical testing conditions of test lengths, sample size, and typical types of non-normal distributions likely to be realized in practice. Results of this dissertation clearly indicate the impact of latent non-normality on estimates of item parameters from dichotomous MixIRT models. Additional manipulation of these conditions might be useful to consider for studying the impact of different priors in the estimation of model parameters.

The objective of this dissertation was to explore the impact of different ability distribution conditions on extraction of spurious latent classes for dichotomous MixIRT models. It would be useful to extend results of this study to determine whether more complex MixIRT models, such as polytomous MixIRT models and multilevel MixIRT models are also susceptible to the kind of over-extraction problem observed here. It would be useful to conduct MixIRT analyses with polytomous type data sets in order to examine the effect of non-normal ability distribution. It also would be useful to explore the effects of non-normality using other estimation techniques such as maximum likelihood and with different MixIRT models, such as mixture random item models.

Bibliography

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723. doi:10.1109/TAC.1974.1100705

[2] Albert, J. H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Educational Statistics, 17*, 251–269. doi:10.3102/10769986017003251

[3] Alexeev, N., Templin, J., & Cohen, A. S. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement, 48*, 313–332. doi:10.1111/j.1745-3984.2011.00146.x

[4] Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of a latent population distribution. *Psychometrika, 42*, 357–374. doi:10.1007/BF02293656

[5] Arminger, G., Stein, P., & Wittenberg, J. (1999). Mixtures of conditional mean- and covariance-structure models. *Psychometrika, 64*, 475–494. doi:10.1007/BF02294568

[6] Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling approach. *Applied Psychological Measurement, 22*, 153–169. doi:10.1177/01466216980222005

[7] Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

[8] Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research, 42*, 757–786. doi:10.1080/00273170701710338

[9] Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for over-extraction of latent trajectory classes. *Psychological Methods, 8*, 338–363. doi:10.1037/1082-989X.8.3.338

[10] Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods, 9*, 3–29. doi:10.1037/1082-989X.9.1.3

[11] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

[12] Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459. doi:10.1007/BF02293801

[13] Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261–280. doi:10.1177/014662168801200305

[14] Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*, 179–197. doi:10.1007/BF02291262

[15] Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple choice data. *Journal of Educational and Behavioral Statistics, 26*, 381–410. doi:10.3102/10769986026004381

[16] Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331–348. doi:10.1111/j.1745-3984.2002.tb01146.x

[17] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345–370.

[18] Cheng, R. C. H., & Traylor, L. (1995). Non-regular maximum likelihood problems. *Journal of the Royal Statistical Society, Series B, 57*, 3–44. Retrieved from http://www.jstor.org/stable/2346086

[19] Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2013). Markov chain Monte Carlo estimation of a mixture item response model. *Journal of Statistical Computation and Simulation, 83*, 278–306. doi:10.1080/00949655.2011.603090

[20] Cho, S.-J., Cohen, A. S., Kim, S.-H., & Bottge, B. (2010). Latent transition analysis with a mixture item response theory measurement model. *Applied Psychological Measurement, 34*, 483–504. doi:10.1177/0146621610362978

[21] Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). New York, NY: Plenum Press.

[22] Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*, 133–148. doi:10.1111/j.1745-3984.2005.00007

[23] Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research & Practice, 20*, 225–233. doi:10.1111/j.1540-5826.2005.00138.x

[24] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

[25] Cohen, J., Cohen, P., West, S. G., & Aiken, L. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

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

[27] Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91*, 883–904. doi:10.1080/01621459.1996.10476956

[28] Dai, Y. (2009). *A mixture Rasch model with a covariate: A simulation study via Bayesian Markov chain Monte Carlo estimation* (Unpublished doctoral dissertation). University of Maryland, College Park.

[29] De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

[30] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1–38. Retrieved from http://www.jstor.org/stable/2984875

[31] Du Toit, M. (Ed.) (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL : Scientific Software International.

[32] Eid, M., & Zickar, M. J. (2007). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 255–270). New York, NY: Springer.

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

[34] Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions.* London, England: Chapman & Hall.

[35] Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521–532. doi:10.1007/BF02293811

[36] Follman, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika, 53*, 553–562. doi:10.1007/BF02294407

[37] Foy, P., Arora, A., & Stanco, G. M. (2013). *TIMSS 2011 user guide for the international database.* Chestnut Hill, MA: Boston College.

[38] Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267–269. doi:10.1207/s15327906mbr23029

[39] Frick, H., Strobl, C., Leisch, F., & Zeileis, A. (2012). Flexible Rasch mixture models with package psychomix. *Journal of Statistical Software, 48*(7), 1–25. Retrieved from http://www.jstatsoft.org/v48/i07/

[40] Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science, 7*, 457–472. Retrieved from http://www.jstor.org/stable/2246093

[41] Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for Gibbs' sampling. *Applied Statistician, 41*, 337–348. Retrieved from http://www.jstor.org/stable/2347565

[42] Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach.* Boca Raton, FL: CRC press.

[43] Gregoire, T. G., & Driver, B. L. (1987). Analysis of ordinal data to detect population differences. *Psychological Bulletin, 101*, 159–165. doi:10.1037/0033-2909.101.1.159

[44] Hannan, E. J. (1987). Rational transfer function approximation. *Statistical Science, 2*, 135–161.

[45] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*, 97–109. doi:10.1093/biomet/57.1.97

[46] Howell, D. C. (2014). *Fundamental statistics for the behavioral sciences*. Belmont, CA: Cengage Learning.

[47] Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science, 16*, 39–59. doi:10.1287/mksc.16.1.39

[48] Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York, NY: Springer.

[49] Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively relevant assessment*. Retrieved from Carnegie Mellon University, Statistics Department, http://www.stat.cmu.edu/brian/nrc/cfa/documents /final.pdf

[50] Kirisci, L., Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*, 146–162. doi:10.1177/01466210122031975

[51] Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26,* 457–477. doi:10.1207/s15327906mbr2603_5

[52] Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and practices* (2nd ed.). New York, NY: Springer.

[53] Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. New York, NY: Houghton-Mifflin.

[54] Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28*, 612–625. doi:10.1111/j.1468-2958.2002.tb00828.x

[55] Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*, 353–373. doi:10.1177/0146621608326422

[56] Li, F., Duncan, T. E., & Duncan, S. C. (2001). Latent growth modeling of longitudinal data: A finite growth mixture modeling approach. *Structural Equation Modeling, 8*, 493–530. doi:10.1207/S15328007SEM0804_01

[57] Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767–778. doi:10.1093/biomet/88.3.767

[58] Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

[59] Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*, 21–39. doi:10.1037/1082-989X.10.1.21

[60] McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6,* 379–396. doi:10.1177/014662168200600402

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

[62] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087–1091. http://dx.doi.org/10.1063/1.1699114

[63] Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–66.

[64] Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359–381. doi:10.1007/BF02306026

[65] Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*, 3–31.

[66] Mislevy, R. J., & Bock, R. D. (1990). BILOG-3: Item analysis and test scoring with binary logistic models [Computer software]. Chicago: Scientific Software International.

[67] Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195–215. doi:10.1007/BF02295283

[68] Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43,* 551–560. doi:10.1007/BF02293813

[69] Muthén, B. (2008). Latent variable hybrids: Overview of old and new models. In G. R. Hancock, & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 1–24). Charlotte, NC: Information Age Publishing.

[70] Muthén, L. K., & Muthén, B. O. (2011). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.

[71] Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18,* 41–68. doi:10.3102/10769986018001041

[72] Nandakumar, R., & Yu, F. (1996). Empirical validation of DIMTEST on non-normal ability distributions. *Journal of Educational Measurement, 33*, 355–368. doi: 10.1111/j.1745-3984.1996.tb00497.x

[73] Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics, 8*, 343–366. Retrieved from http://www.jstor.org/stable/2369392

[74] Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

[75] Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14*, 535–569. doi:10.1080/10705510701575396

[76] Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146–178. doi:10.3102/10769986024002146

[77] Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philisophical Transactions of the Royal Society of London. A, 185*, 71–110. Retrieved from http://www.jstor.org/stable/90667

[78] Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika, 62*, 223–241. doi:10.1093/biomet/62.2.223

[79] Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News, 6*, 7–11. Retrieved from http://cran.r-project.org/doc/Rnews/Rnews_2006-1.pdf#page=7

[80] Preinerstorfer, D. (2011). mRm: An R package for conditional maximum likelihood estimation in mixed Rasch models. Retrieved from: http://cran.r-project.org/web/packages/mRm/ index.html

[81] Preinerstorfer, D., & Formann, A. K. (2011). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology, 65*, 251–262. doi:10.1111/j.2044-8317.2011.02020.x

[82] R Development Core Team (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

[83] Raftery, A. E., & Lewis, S. (1992). How many iterations in the Gibbs sampler. *Bayesian statistics, 4*, 763–773.

[84] Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielsen and Lydiche.

[85] Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207–230. doi:10.3102/10769986004003207

[86] Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

[87] Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271–282. doi:10.1177/014662169001400305

[88] Rost, J. (1991). A logistic mixture distribution model for polytomous item response. *British Journal of Mathematical and Statistical Psychology, 44*, 75–92.

[89] Rost, J., & von Davier, M. (1993). Measuring different traits in different populations with the same items. In R. Steyer, K. F. Wender, & K. F. Widaman (Eds.), *Psychometric methodology. Proceedings of the 7th European Meeting of the Psychometric Society in Trier* (pp. 446–450). Stuttgart, Germany: Gustav Fischer Verlag.

[90] Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In I. W. Molenaar (Eds.), *Rasch models-foundations, recent developments and applications* (pp. 257–268). New York, NY: Springer Verlag.

[91] Rost, J., Carstensen, C. H., & von Davier, M. (1997). Applying the mixed-Rasch model to personality questionaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324–332). New York, NY: Waxmann.

[92] Samuelsen, K. M. (2005). *Examining differential item functioning from a latent class perspective* (Doctoral dissertation). University of Maryland, College Park.

[93] SAS Institute. (2008). *SAS/STAT 9.2 user's guide*. Cary, NC: Author.

[94] Sass, D. A., Schmitt, T. A., & Walker, C. M. (2008). Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education, 21*, 65–88. doi:10.1080/08957340701796415

[95] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464. doi:10.1214/aos/1176344136

[96] Sclove, L. S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333–343. doi:10.1007/BF02294360

[97] Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*, 299–311. doi:doi: 10.1177/014662169001400307

[98] Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*, 375–394. doi:10.1111/j.1745-3984.2005.00021.x

[99] Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*, 298–321. doi:10.1177/0146621605285517

[100] Smith, B. J. (2007). boa: An R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software, 21*, 1–37. Retrieved from http://www.jstatsoft.org/v21/i11/paper

[101] Spiegelhalter, D. J., Best, N. G., & Carlin, B. P. (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models* (Research Report No. 98-009). MRC Biostatistics Unit, Cambridge.

[102] Spiegelhalter, D., Thomas, A., & Best, N. (2003). WinBUGS (version 1.4) [Computer software]. Cambridge, UK: Biostatistics Unit, Institute of Public Health.

[103] Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika, 55*, 461–475. doi:10.1007/BF02294761

[104] Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion referenced test. *Journal of Educational Measurement, 13*, 265–276. doi:10.1111/j.1745-3984.1976.tb00017.x

[105] Thissen, D. (1991). *MULTILOG users guide: Multiple categorical item analysis and test scoring using item response theory.*: Chicago, IL: Scientific Software International.

[106] Thissen, D. (2003). MULTILOG: multiple, categorical item analysis and test scoring using item response theory (Version 7.03) [Computer software]. Chicago, IL: Scientific Software International.

[107] TIMSS 2011 Assessment. Copyright 2012 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA and International Association for the Evaluation of Educational Achievement (IEA), IEA Secretariat, Amsterdam, the Netherlands.

[108] Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester, England: Wiley.

[109] Tofighi, D., & Enders, C. K. (2007). Identifying the correct number of classes in a growth mixture model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Mixture models in latent variable research* (pp. 317–341). Greenwich, CT: Information Age.

[110] van den Oord, E. J. C. G. (2005). Estimating Johnson curve population distributions in MULTILOG. *Applied Psychological Measurement, 29*, 45–64. doi:10.1177/0146621604269791

[111] Vermunt, J. K., & Magidson, J. (2005). Latent GOLD (Version 4.0) [Computer software]. Belmont, MA: Statistical Innovations, Inc.

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

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

[114] von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 371–379). New York, NY: Springer.

[115] von Davier, M., & Rost, J. (1997). Self monitoring-A class variable? In J. Rost & R. Langeheime (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 296–305). Muenster, Germany: Waxmann.

[116] von Davier, M., & Rost, J. (2007). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 643–661). Amsterdam, The Netherlands: Elsevier.

[117] von Davier, M., Rost, J., & Carstensen, C. H. (2007). Introduction: Extending the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 1–12). New York, NY: Springer.

[118] von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measurement, 28*, 389–406. doi:10.1177/0146621604268734

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

[120] Wall, M. M., Guo, J., & Amemiya, Y. (2012). Mixture factor analysis for approximating a nonnormally distributed continuous latent factor with continuous and

dichotomous observed variables. *Multivariate Behavioral Research, 47*, 276–313. doi:10.1080/00273171.2012.658339

[121] Willse, J. T. (2009). mixRasch: Mixture Rasch models with JMLE (Version 0.1). Retrieved from http://CRAN.R-project.org/package=mixRasch

[122] Willse, J. T. (2011). Mixture Rasch models with joint maximum likelihood estimation. *Educational and Psychological Measurement, 71(1)*, 5–19. doi:10.1177/0013164410387335

[123] Wilson, D. T., Wood, R., & Gibbons, R. (2003). TESTFACT: Test scoring, item statistics, and item factor analysis [Computer software]. Chicago, IL: Scientific Software International.

[124] Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8*, 347–364. doi:10.1177/014662168400800312

[125] Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement, 40*, 307–330. doi:10.1111/j.1745-3984.2003.tb01149.x

[126] Woods, C. M. (2004). *Item response theory with estimation of the latent population distribution using spline-based densities* (Unpublished doctoral dissertation). University of North Carolina at Chapel Hill.

[127] Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods, 11*, 253–270. doi:10.1037/1082-989X.11.3.253

[128] Woods, C. M. (2006b). *RCLOG v.2: Software for item response theory parameter estimation with the latent population distribution represented using spline-based densities* (Tech. Rep.). St. Louis: Washington University.

[129] Woods, C. M. (2007). Ramsay curve IRT for Likert-type data. *Applied Psychological Measurement, 31*, 195–212. doi:10.1177/0146621606291567

[130] Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement, 32*, 511–526. doi:10.1177/0146621607310402

[131] Woods, C. M., & Thissen, D. (2004). *RCLOG v.1: Software for item response theory parameter estimation with the latent population distribution represented using spline-based densities* (Tech. Rep.). Chapel Hill: L. L. Thurstone Psychometric Laboratory, University of North Carolina.

[132] Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika, 71*, 281–301. doi:10.1007/s11336-004-1175-8

[133] Xu, X., & Jia, Y. (2011). *The sensitivity of parameter estimates to the latent ability distribution* (ETS RR-11-40). Retrieved from Educational Testing Service website: http://www.ets.org/Media/Research/pdf/RR-11-40.pdf

[134] Yamamoto, K. Y., & Everson, H. T. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). Munster, Germany: Waxmann.

[135] Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213–249. doi:10.1007/BF02294536

[136] Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). BILOG-MG 3 [Computer software]. Lincolnwood, IL: Scientific Software International.

[137] Zwinderman, A. H., & Van den Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Applied Psychological Measurement, 14*, 73–81. doi:10.1177/014662169001400107

Appendix A

Gelman-Rubin Shrink Factor Plots

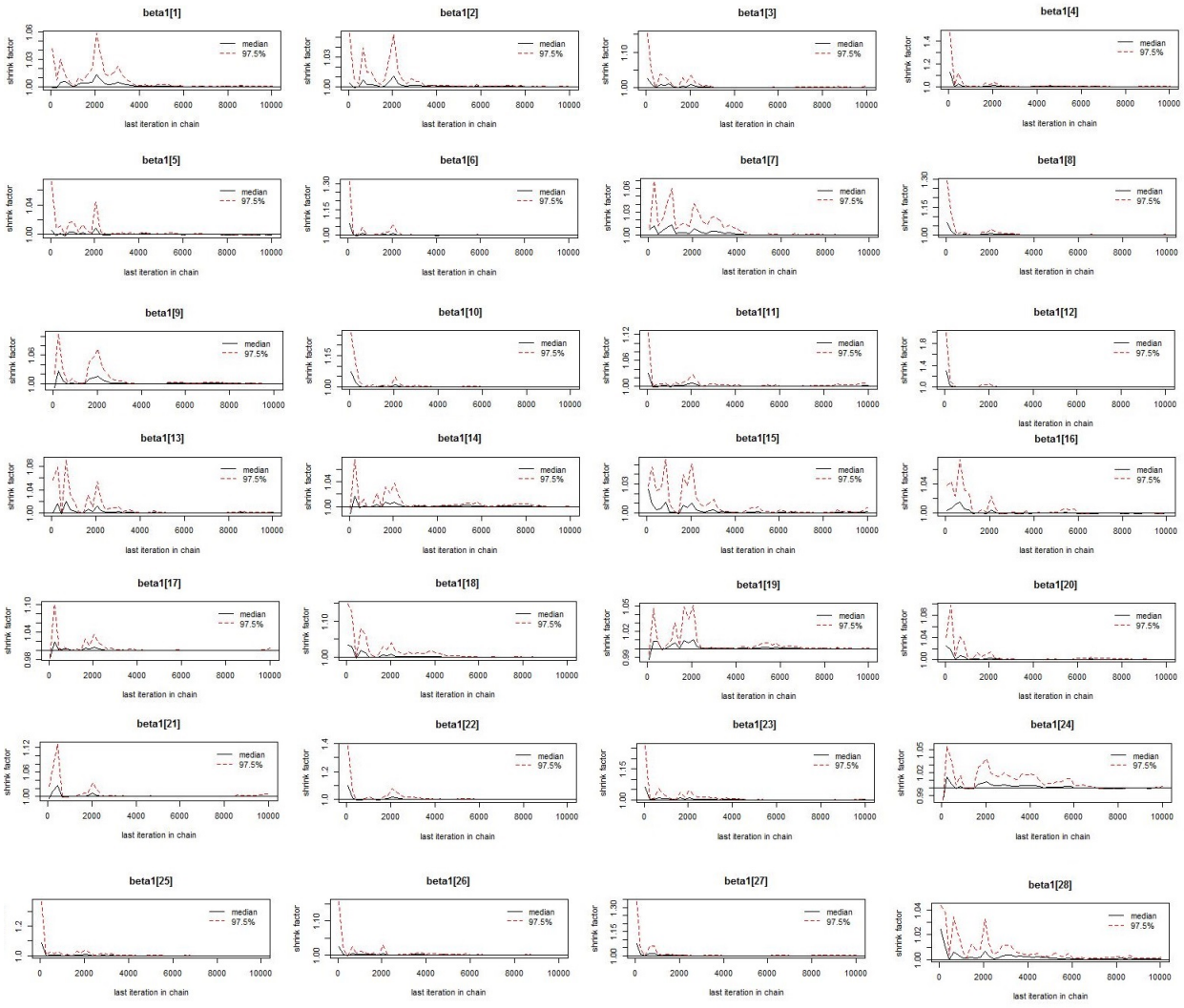

Figure A.1: Gelman and Rubin Shrink Factor Plots for MRM with 28 Items

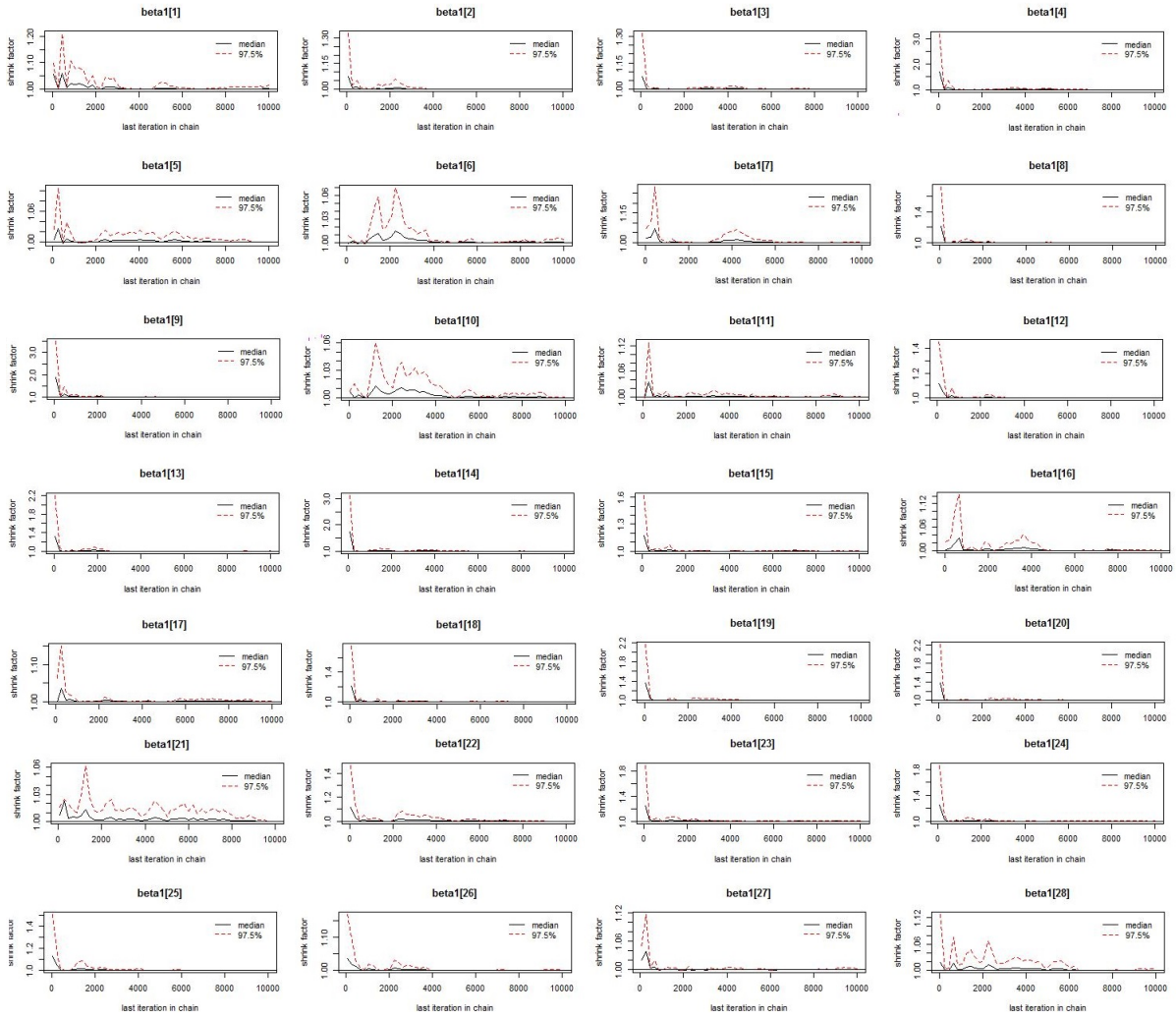

Figure A.2: Gelman and Rubin Shrink Factor Plots for Mix2PL IRT Model with 28 Items

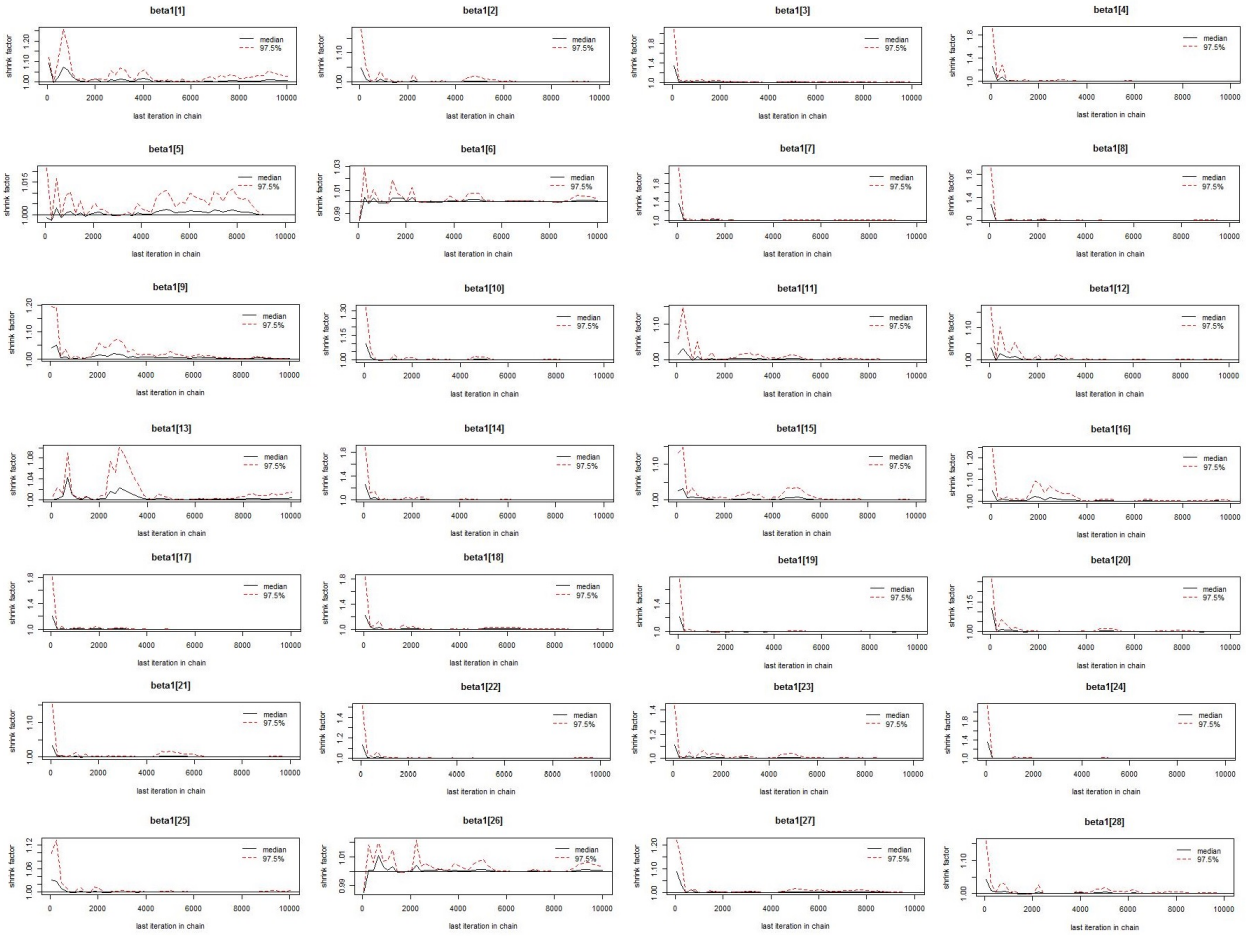

Figure A.3: Gelman and Rubin Shrink Factor Plots for Mix3PL IRT Model with 28 Items

Appendix B

WinBUGS Codes for Data Analyses

```
#################
### MRM #########
#################
model
{
  for (j in 1:NE) {
   for (k in 1:NI) {
          r1[j,k]<-resp[j,k]
          r2[j,k]<-resp[j,k]

}}
#  1group model
for (j in 1:NE) {
for (k in 1:NI) {
     tt1[j,k]<- exp(theta1[j] - beta1[k])
     p1[j,k]<-tt1[j,k]/(1 + tt1[j,k])
     r1[j,k]~dbern(p1[j,k])
     l1[j,k]<-log(p1[j,k])*r1[j,k]+log(1-p1[j,k])*(1-r1[j,k])

}
```

```
}
     loglik[1]<-sum(l1[1:NE,1:NI])

#  Priors for 1group
for (j in 1:NE) {
theta1[j] ~ dnorm(0, 1)
}
 for (k in 1:NI) {
    beta1[k]~dnorm(0,1)
 }



# 2group model
   for (j in 1:NE) {
       for (k in 1:NI) {
           tt2[j,k]<- exp(theta2[j] - beta2[gmem2[j],k])
           p2[j,k]<-tt2[j,k]/(1 + tt2[j,k])
           r2[j,k]~dbern(p2[j,k])
          l2[j,k]<-log(p2[j,k])*r2[j,k]+log(1-p2[j,k])*(1-r2[j,k])

           }
         theta2[j] ~ dnorm(mut2[gmem2[j]],1)
         gmem2[j] ~ dcat(pi2[1:G2])
        }
       loglik[2]<-sum(l2[1:NE,1:NI])
```

```
# Priors for 2group
pi2[1:G2]~ ddirch(alpha2[])
for (j in 1:G2){
    for (k in 1:NI){
          beta2[j,k]~dnorm(0,1)
         }
     }
for (j in 1:G2) {
      mut2[j]~ dnorm(0.,1.)
      }

for (j in 1:2){
par[j]<-NI*j+2*j-1
AIC[j]<--2*loglik[j]+2*par[j]
BIC[j]<--2*loglik[j]+par[j]*log(NE)
}}

###################
### Mix2PL MODEL###
###################
 model
{
  for (j in 1:NE) {
   for (k in 1:NI) {
          r1[j,k]<-resp[j,k]
          r2[j,k]<-resp[j,k]
```

```
}}



#  1group model
for (j in 1:NE) {
for (k in 1:NI) {
     tt1[j,k]<- exp(a1[k]*(theta1[j] - beta1[k]))
     p1[j,k]<-tt1[j,k]/(1 + tt1[j,k])
     r1[j,k]~dbern(p1[j,k])
     l1[j,k]<-log(p1[j,k])*r1[j,k]+log(1-p1[j,k])*(1-r1[j,k])

}
}
loglik[1]<-sum(l1[1:NE,1:NI])

#  Priors for 1group
for (j in 1:NE) {
theta1[j] ~ dnorm(0, 1)
}
 for (k in 1:NI) {
    a1[k]~dnorm(0,1)I(0,)
    beta1[k]~dnorm(0,1)
 }

# 2group model
   for (j in 1:NE) {
```

```
        for (k in 1:NI) {
            tt2[j,k]<- exp(a2[gmem2[j],k]*(theta2[j] - beta2[gmem2[j],k]))
            p2[j,k]<-tt2[j,k]/(1 + tt2[j,k])
            r2[j,k]~dbern(p2[j,k])
            l2[j,k]<-log(p2[j,k])*r2[j,k]+log(1-p2[j,k])*(1-r2[j,k])
            }
          theta2[j] ~ dnorm(mut2[gmem2[j]],1)
          gmem2[j] ~ dcat(pi2[1:G2])
         }
loglik[2]<-sum(l2[1:NE,1:NI])

# Priors for 2group
pi2[1:G2]~ ddirch(alpha2[])
for (j in 1:G2){
    for (k in 1:NI){
          beta2[j,k]~dnorm(0,1)
          a2[j,k]~dnorm(0,1)I(0,)
    }
}
for (j in 1:G2){
      mut2[j]~ dnorm(0.,1.)
}

for (j in 1:5){
par[j]<-2*NI*j+2*j-1
AIC[j]<--2*loglik[j]+2*par[j]
```

```
BIC[j]<--2*loglik[j]+par[j]*log(NE)
 }}

###################
### Mix3PL MODEL###
###################
 model
{
  for (j in 1:NE) {
   for (i in 1:NI) {
          r1[j,i]<-resp[j,i]
          r2[j,i]<-resp[j,i]

}}

#  1-group 3PL model
for (j in 1:NE) {
for (i in 1:NI) {
   tt1[j,i]<- exp(a1[i]*(theta1[j] - beta1[i]))
   logit1[j,i]<-tt1[j,i]/(1 + tt1[j,i])
   p1[j,i]<-c1[i]+(1-c1[i])*logit1[j,i]
   r1[j,i]~dbern(p1[j,i])
   l1[j,i]<-log(p1[j,i])*r1[j,i]+log(1-p1[j,i])*(1-r1[j,i])

}}

#  Priors for 1-group 3PL model
```

```
for (j in 1:NE) {
    theta1[j] ~ dnorm(0, 1)
}
 for (i in 1:NI) {
    a1[i]~dnorm(0,1)I(0,)
    beta1[i]~dnorm(0,1)
    c1[i]~dbeta(5.,17.)
}

# 2-group 3PL model
   for (j in 1:NE) {
       for (i in 1:NI) {
            tt2[j,i]<- exp(a2[gmem2[j],i]*(theta2[j] - beta2[gmem2[j],i]))
            logit2[j,i]<-tt2[j,i]/(1 + tt2[j,i])
            p2[j,i]<-c2[gmem2[j],i]+ (1-c2[gmem2[j],i])*logit2[j,i]
            r2[j,i]~dbern(p2[j,i])
            l2[j,i]<-log(p2[j,i])*r2[j,i]+log(1-p2[j,i])*(1-r2[j,i])

           }
           theta2[j] ~ dnorm(mut2[gmem2[j]],1)
           gmem2[j] ~ dcat(pi2[1:G2])
        }

# Priors for 2-group 3PL model
pi2[1:G2]~ ddirch(alpha2[])
for (k in 1:G2){
```

```
    for (i in 1:NI){
          beta2[k,i]~dnorm(0,1)
          a2[k,i]~dnorm(0,1)I(0,)
          c2[k,i]~dbeta(5., 17.)
         }
     }
#mut2[1]<-0
for (k in 1:G2) {
      mut2[k]~ dnorm(0.,1)
      }

loglii[1]<-sum(l1[1:NE,1:NI])
loglii[2]<-sum(l2[1:NE,1:NI])

for(k in 1:2){
par[k]<-3*NI*k+2*k-1
AIC[k]<--2*loglii[k]+2*par[k]
BIC[k]<--2*loglii[k]+par[k]*log(NE)
}}
```