This is to certify that the
dissertation entitled

INVESTIGATING UNOBSERVED HETEROGENEITY USING ITEM RESPONSE THEORY MIXTURE MODELS

presented by

DIPENDRA RAJ SUBEDI

has been accepted towards fulfillment
of the requirements for the

Ph. D. degree in Measurement and Quantitative Methods

Mark D. Reckase
Major Professor's Signature

July 10, 2009
Date

INVESTIGATING UNOBSERVED HETEROGENEITY USING ITEM RESPONSE THEORY MIXTURE MODELS

By

Dipendra Raj Subedi

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods

2009

ABSTRACT

INVESTIGATING UNOBSERVED HETEROGENEITY USING ITEM RESPONSE THEORY MIXTURE MODELS

By

Dipendra Raj Subedi

Many item response theory (IRT) scaling and scoring models assume that examinee samples have comparable test-taking behaviors and comparable performance among different subgroups (e.g., gender and ethnicity) or in different test-taking contexts (e.g., geographic location or test-taking mode). However, in some situations the aforementioned assumption of test-taking homogeneity may not hold and test-taking heterogeneity is said to exist. When these sources of heterogeneity are unobservable (e.g., when examinees have unexpected guessing behaviors), then IRT mixture modeling (MixIRT) may be preferable to traditional IRT (i.e., two- and three-parameter logistic models) modeling for adjusting the parameter estimation inaccuracies that might otherwise occur in the presence of unobserved heterogeneity.

Therefore, the goals of this study were to investigate: a) the estimation accuracy of MixIRT models when test-taking heterogeneity exists and b) the efficiency of MixIRT models in identifying subsets of examinees whose item responses do not fit the specified IRT model. Additionally, given the difficulty in estimating MixIRT parameters, Bayesian modeling with the Markov chain Monte Carlo method was used and the robustness of

MixIRT modeling was investigated through a simulation study. This simulation study investigated several realistic testing factors that included test-taker sample size, test length, and the proportion of test-taking heterogeneity in the form of examinee guessing behavior. In other words, varying these testing factors allowed the evaluation of the impact of test-taking heterogeneity on the accuracy of parameter estimation. The results of the simulation study showed that the MixIRT model provided more accurate parameter estimates than traditional IRT models and was quite efficient in identifying subsets of examinees that had anomalous test-taking behaviors. A real data example also corroborated the simulation study results.

Dedication

To my beloved parents: Bhoj Raj Subedi and Shanti Subedi

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to many people who assisted me throughout my doctoral studies in the Measurement and Quantitative Methods program at Michigan State University.

First of all, I would like to express my sincere gratitude to my dissertation director and co-chair of the guidance committee, Dr. Mark Reckase. I am deeply indebted to him for his guidance and tremendous support throughout my dissertation. Dr. Kimberly Maier, co-chair of the guidance committee provided continuous support from the beginning of my doctoral studies. Other dissertation committee members Dr. Yeow Meng Thum and Dr. Punya Mishra also provided excellent help and insightful comments during the various stages of my dissertation.

My special appreciation goes to Dr. Raymond Mapuranga for his excellent help and camaraderie from the beginning of my doctoral studies to the completion of this dissertation. I especially appreciate his editorial comments in my dissertation. I would also like to thank Dr. Joseph Martineau at Michigan Department of Education for giving me an excellent opportunity to gain hands-on experience on various psychometric analyses while I was still a graduate student. I am very grateful to Professor Murari Suvedi and Dr. Bishwa Adhikari for their continuous support from the beginning of my graduate studies at MSU.

I would like to express my special thanks to the Graduate School at MSU for providing me a dissertation completion fellowship. Thanks to my writing group members at MSU: Michael Sherry, Sungworn Ngudgratoke, and Young Yee Kim for their constructive feedbacks and comments in my writing. Also, thanks to my professors at MSU, who provided me various teaching and research assistantships. Thanks to Adam Wyse and Minh Duong for their wonderful friendship. I would also like to take this opportunity to express my appreciation to my colleagues and mentors at the American Institutes for Research, Washington D.C. for their understanding during the final editing stage of my dissertation.

Finally, my special thanks go to my family. First, to my wife Shanta Subedi, for her love, patience, and continuous support throughout this journey. To my parents, who stood at every stage of my career with their unconditional support. To my brother Bishwa Subedi and other family members for their wonderful support.

TABLE OF CONTENTS

LIST OF TABLES

4.13

4.14

4.15

4.16

4.17

4.18

4.19

4.20

4.21

4.22

LIST OF FIGURES

KEY TO ABBREVIATIONS

IRT	Item response theory
2PL	Two-parameter logistic model
3PL	Three-parameter logistic model
MixIRT	Mixture item response theory model
MIRT	Multidimensional item response theory model
RMSE	Root mean squared error
MixIRT-R	Mixture item response theory model with random guessing
MixIRT-A	Mixture item response theory model with ability-based guessing

CHAPTER 1
INTRODUCTION

1.1 Background

The K-12 test-based accountability system has gained increasing attention from researchers, educators, and policy-makers since the implementation of the "No Child Left Behind" legislation (NCLB, 2001). This legislation was designed to improve existing educational practice and student academic achievement through improved teaching and curriculum (Hamilton, Stecher, & Klein, 2002). This renewed attention to K-12 education also led to increased interest in comparative and international assessments because of the ever-widening subject matter knowledge gap between US students and their counterparts in other industrialized countries. For example, results from the Third International Mathematics and Science Survey showed that US fourth and eighth grade students had comparatively lower mathematics and science achievement than many other developed countries (Gonzales et al., 2000; Lemke & Gonzales, 2006).

This increased focus on large scale assessment has also resulted in the added scrutiny of psychometric modeling approaches. Specifically, the accuracy of parameter estimates when using traditional psychometric models has come into question because they do not efficiently account for anomalous examinee behaviors such as cheating and guessing. These undesirable examinee behaviors can lead to aberrant responses that can occlude the accuracy of inferences drawn about student knowledge and academic skills.

1.2 Anomalous Examinee Test-taking Behavior

As noted previously, the accuracy of psychometric parameter estimates used in large-scale assessments is fallible when examinees exhibit anomalous test-taking behavior. Common anomalous behaviors include guessing, cheating and examinee motivation. Cheating can be defined as any action that decreases the accuracy of the intended inferences based on the examinee's performance, thus threatening the validity of the inference about the test taker (Cizek, 2001). Examinee motivation can be defined as the degrees of effort test-taker expend particularly when given a low-stakes assessment test. Several recent studies have investigated each of these testing phenomena. However, this study focuses only on heterogeneity introduced by test-takers' guessing behavior.

Guessing may occurs when test-takers run out of time on the test, when they are less motivated, or when they find test items difficult. Guessing behavior, however, varies depending upon the nature of the test (low-stakes or high-stakes), item difficulty, examinee ability, available time to complete the test, and cross-cultural differences among examinees.

The validity of the inference made using scores is partially dependent on the amount of effort put forth by the examinee while taking the test (Wise, 2006). Furthermore, when adequate effort is not given by examinees, they tend to guess randomly which makes it difficult to estimate the test taker's true subject matter proficiency (Budescu & Bar-Hillel, 1993).

Therefore, anomalous test-taking behavior is a concern and is particularly a concern in low-stakes tests when examinees are more likely to have low motivation, and to cheat or guess excessively. In low-stakes tests like the National Assessment of

Educational Progress, attempts to mitigate the negative consequences of unusual examinee behavior include the use of shorter tests with specialized data collection designs like balanced incomplete blocking (Johnson, 1992). Even formula scoring does not adequately deter guessing on tests (Frary, 1988).

It is also particularly important to identify and account for anomalous examined behavior in the current NCLB era because examinee test scores are an integral aspect of data-driven educational policy. For example, AYP decisions are based strongly on examinee test scores and yet not many model-based or sample-specific adjustments are made in the estimation of psychometric parameters.

1.3 Test-taking Heterogeneity

It is important to study the undesirable test-taking behaviors described previously because they can have negative consequences on the interpretation and accuracy of psychometric models. Particularly, the psychometric models used in low-stakes tests have the underlying assumption that the same item parameters and ability distributions apply to all examinees taking the test. This is known as the assumption of test-taking homogeneity (e.g., Baker & Kim, 2004; Bock & Zimowski, 1997; Lord, 1980). But as noted previously, examinees often exhibit unconventional test-taking patterns such as cheating, excessive guessing and low motivation.

When the aforementioned anomalous behaviors exist, test-taking heterogeneity is said to exist and is often evidenced by sample variability that occurs among different groups of test-takers. An example pertaining to this study would be guessers and non-guessers. Similarly, test-taking behavior may be different for different groups of

examinees. For example, the Graduate Record Examinations' (GRE) verbal assessment is administered to native and non-native English speakers who have different English proficiency which could impact their performance on the test and not allow the two groups to be analyzed together.

In situations where the group-membership of test-takers is not observable, *unobservable test-taking heterogeneity* is said to exist. In contrast, if the source of heterogeneity can be observed in the data (e.g., gender, ethnicity), *observable heterogeneity* makes it convenient to stratify test-takers for any validation using multi-group analyses (Muthén & Lehman, 1985).

Multi-group analyses are important in psychometrics and when test-taker characteristics are not observable, a set of models called *latent class models* can be used for multi-group analyses. In the case of low-stakes tests, a form of latent class models called *mixture models* have been used recently by researchers for multi-group analyses when test-taking heterogeneity is unobservable (Bock & Zimowski, 1997). As a result, these mixture models were extended to models commonly used in low-stakes tests under a framework which analyses the interaction between examinee ability and test items called *item response theory* (Lord, 1980). IRT is the most common modeling approach used in many tests like NAEP, TIMSS and PISA.

1.4 Traditional Item Response Theory Modeling

Traditional item response theory (IRT) modeling allows examinee performance on each test item to be succinctly quantified across all examinees. The three assumptions of traditional IRT are: dimensionality, local independence, and the existence of a monotonically increasing function (Hambleton, 1989). First, the dimensionality

assumption indicates that a test should measure only one ability, personality trait or attitude -- called *unidimensionality*. When more than one ability is assumed to exist, these IRT models are called *multidimensional* (Hambleton & Swaminathan, 1985; Reckase, 1997). Local independence implies that no item should provide clues to the answers of other items in a test (Hambleton & Swaminathan, 1985). Finally, the assumption of a monotonically increasing function relates the probability of success on an item to the ability measured by the item.

A common traditional IRT model is the three-parameter logistic (3PL) model which is represented mathematically as:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}, \quad i = 1, 2, ..., n \qquad (1.1)$$

where $P_i(\theta)$ is the probability that a given test-taker with ability θ answer a random item correctly. a_i is the item discrimination, b_i is the item difficulty and c_i is the pseudo guessing parameter (Hambleton & Swaminathan, 1985). Another common model called the two-parameter logistic (2PL) model is obtained when $c = 0$ in Equation 1.1 and the one-parameter logistic (1PL) model is obtained when $c = 0$ and $a = 1$.

Although these traditional IRT models are useful for quantifying examinee ability, they are not able to account for unobservable test-taking heterogeneity, which may result in parameter estimates that are inaccurate. As noted above, mixture models are capable of accounting for unobservable heterogeneity and extensions of these models in IRT framework have produced so-called mixture IRT models or MixIRT for short.

Therefore, MixIRT models provide greater flexibility in modeling complex item response distributions (McLachlan & Peel, 2000). Hence, in the low-stakes testing context, MixIRT models would be particularly useful and provide impetus for investigation of their robustness in modeling anomalous test-taking behavior as described in the section which follows.

1.5 Motivation

Given the psychometric modeling limitations of traditional IRT models in accounting for test-taking heterogeneity, it is important to investigate the efficiency and accuracy of MixIRT models in estimating parameters. Moreover, the estimation of latent distributions (e.g., Mislevy, 1984) is an important area of psychometric research because even the most intuitively appealing and creative models are not useful unless the parameters in the model can be estimated accurately. Specifically, modeling unobserved test-taking heterogeneity such as aberrant item responses is crucial because ignoring it can lead to biased parameter estimates and may yield inflated measurement and test reliability (Lord & Novick, 1968; Muthén, 1989). Furthermore, the inaccurate estimation of examinee latent traits can have consequential impacts such as false interpretation of student ability, and erroneous measurement of school and teacher effectiveness (Ansari, Jedidi, & Dube, 2002).

Given the limitations of IRT modeling articulated above, specifically in the modeling of guessing, the 3-PL model -- a commonly used model -- is unlikely to suffice for psychometric modeling in large-scale assessments. This is because it restricts the guessing parameter to be item dependent. Most importantly, the 3PL model is incapable of identifying whether individual test-takers actually guess, but rather it models guessing

over the entire sample and hence results in inadequate modeling of guessing. Therefore, a subsidiary motivation of this study is to explicate the implications of inaccurately modeling guessing or random response behavior when this phenomenon is not modeled at the person level, but rather at the item level as is in current IRT modeling practice. Moreover, the MixIRT approach taken in this study is more appropriate than IRT for providing evidence of the impact of guessing on individual test items and has the secondary advantage of possibly identifying students with low motivation.

1.6 Purpose

Mixture model parameters are estimated using either frequentist or Bayesian approaches. As described in Chapter Two, several practical problems arise in the frequentist approach to mixture model parameter estimation (Frühwirth-Schnatter, 2006). On the other hand, Bayesian estimation methods can handle high-dimensional problems and allow exploration of the distributions of parameters, regardless of the distributional forms of the likelihood functions or parameters. In addition, model complexity increases with the increase in number of parameters to be estimated, such as a mixture model, particularly with a large number of mixture components.

Therefore, this study focused on using a Bayesian approach to parameter estimation in mixture IRT models, with specific emphasis on item parameter estimation, test-taker cluster identification, and proficiency level classification. These issues are of increased interest among researchers, policymakers, and educators in the current era of test-based accountability systems. In particular, this study compared the performance of Bayesian mixture IRT modeling to common IRT models in estimating person and item parameters, and identifying aberrant responses and low-motivation test-takers.

The remainder of the dissertation is divided into four chapters. Chapter Two reviews the literature that lays out the important empirical and theoretical foundation for this dissertation. The third chapter presents the methodology and the research design, implementation of Bayesian estimation methods, and mixture model analysis. The results from both simulation and empirical data analysis are presented in Chapter Four. Finally, Chapter Five provides discussion, limitations, suggestions for further research, and summary of results and conclusions.

CHAPTER 2
LITERATURE REVIEW

As noted in the previous chapter, the purpose of this study is to investigate and illustrate the efficacy of using mixture models and a Bayesian approach in estimating item parameters and test-taker ability under the IRT framework. Therefore, the purpose of this chapter is to introduce important Bayesian and mixture modeling concepts that are pertinent to this study. In the sections which follow, descriptions of mixture distributions, mixture model parameter estimation, Bayesian statistical modeling, prior research on psychometric applications of mixture models, and the modeling of guessing behaviors in tests, are provided.

2.1 Modeling Sources of Unobserved Heterogeneity

The latent structure model (Goodman, 1974; Lazersfeld & Henry, 1968) is used to explain underlying, unobservable or latent categorical relationships, and offers an efficient way of uncovering distinct sub-populations, incorporating correlated non-normally distributed outcomes, and classifying individuals into classes. That is, these models can serve as possible elucidations of the observed relationships among a set of manifest variables (Goodman, 1974). Depending upon the nature of variables used in these latent structure models, various types of models can be defined under this framework. Specifically, mixture modeling is categorized as a subset of latent structure models when latent variables that represent subpopulations are used for modeling

population membership. Mixture models in the context of IRT are presented next.

2.2 Mixture Distributions and Mixture IRT Models

Mixture distributions are comprised of a finite or infinite number of components, possibly of different distributional types, that can describe different features of data. A mixture model is a flexible tool for modeling complex data through an appropriate choice of data components to accurately represent the data's true characteristics (McLachlan & Peel, 2000). As a result, mixture models are a valuable tool for analyzing a wide variety of latent trait phenomena.

Mathematically, a mixture model can be represented by the observation of n independent random variables $x_1, x_2, \ldots, x_n$, from a k-component mixture density as denoted by Equation 2.1:

$$f(x_i) = \sum_{j=1}^{k} \pi_j \, f_j(x_i), \quad \mathrm{i} = 1, \ldots, \mathrm{n} \qquad (2.1)$$

where $\pi_j > 0$, j=1, …,k; $\pi_1 + \ldots\ldots + \pi_k = 1$ and $\mathrm{f_j(x)}$, $1 \leq \mathrm{j} \leq \mathrm{k}$, are the component densities of the mixture and $\pi_1, \ldots, \pi_k$ are the mixing proportions. These proportions allowed us to estimate the size of subgroups in the sample.

Mixture IRT (MixIRT) models are a combination of LCA and IRT models (Asparouhov & Muthén, 2008). Their development has been motivated primarily by diverse phenomena that are encountered when modeling data from populations that are potentially non-homogeneous (von Davier & Rost, 2007) such as heterogeneous

population of guessers. LCA is a statistical method used to identify homogeneous groups, or classes, from categorical multivariate data. In addition, MixIRT models are useful in testing for the population invariance of item parameters and ability distribution. Basically, these models are based on the assumption that the population under investigation is composed of two or more latent subpopulations dictated by different degrees of latent traits, each of which responds differentially to psychological tasks and stimuli (Draney, Wilson, Gluck, & Spiel, 2008). One of the most general MixIRT models is the mixed Rasch model (Rost, 1990) in which each examinee is parameterized both by a class membership parameter ($g = 1, \ldots . G$) and a within-class ability parameter (θ_g). The probability of a correct response (U) to the item is represented mathematically as:

$$P(U = 1 \mid g, \theta_g) = \frac{e^{(\theta_g - b_{jg})}}{1 + e^{(\theta_g - b_{jg})}} \qquad (2.2)$$

The psychometric applications of mixture modeling, particularly those relevant to this study, are briefly reviewed in the next section.

2.3 Psychometric Applications of Mixture Modeling

While mixture modeling has been used to detect guessers in large scale assessment, it has also been used in various psychometric applications. One of the earliest applications of mixture modeling in psychometrics is the HYBRID model (Yamamoto, 1989), which was used to detect randomly guessed item responses. Mislevy and Verhelst (1990) described a family of multiple-strategy IRT models that apply when each subject belongs to one of a number of exhaustive and mutually exclusive classes that correspond

to item-solving strategies. Other applications of mixture modeling include modeling item response times with a two-state mixture model (Schnipke & Scrams, 1997) that identified guessers. Furthermore, De Ayala, Kim, Stapleton, and Dayton (2002), Cohen and Bolt (2005) and Samuelsen (2005) used the mixture model approach in differential item functioning analysis.

Recent research in multivariate and mixture distribution Rasch models are presented in von Davier and Carstensen (2007), which focused on an extension of the Rasch model in which certain homogeneity assumptions have been relaxed on both item and population levels. Furthermore, some applications of these extensions in educational or psychological contexts are provided.

To identify individual guessers from item response data, Yamamoto (1989, 1995) used the HYBRID model (Yamamoto, 1987). It was mainly focussed on estimating the effect of test length and time on parameter estimation. Similarly, Wise and DeMars (2006) used the effort-moderated IRT model, which in the presence of guessers performed better than the standard 3PL model in terms of model fit, accuracy of item parameter and test information estimation, and the degree of convergence validity in proficiency estimation.

Recently, Yang (2007) reviewed the methods of identifying guessers and proposed approaches for modeling response times based on the two-state mixture model. A majority of these methods used item response time; hence their scope is limited to computer-based or computer-adaptive testing. Additionally, data on item response time is not available in most large scale assessments and any paper-based assessment. The aforementioned models and modeling approach provided motivation for the simulation

study described herein where estimation of the probability of guessing is based entirely on item response patterns and not on response time, similar to a recent study by Cao and Stokes (2008). Parameter estimation is an important issue, which is described next.

2.4 Mixture IRT Model Parameter Estimation

Estimation of person and item parameters is an important problem encountered in applications of item response theory (Hulin, Lissak, & Drasgow, 1982). Parameter estimation is an important issue because even complex models are not useful unless their parameters can be estimated accurately. For example, in this study, it is very important to classify accurately examinees into *guessers* and *non-guessers*. Originally, Pearson (1894) used the methods of moment to estimate mixture model parameters. Later, Rao (1948) introduced the maximum likelihood estimation approach to estimate the mixture model parameters. Rao's approach follows the frequentist paradigm, but an alternative Bayesian approach was introduced by Lavine and West (1992). An overview of both frequentist and Bayesian approaches is provided next.

2.4.1 Frequentist Approach to Parameter Estimation

The Expectation-Maximization (EM) algorithm has been typically used in the estimation of mixture distributions. The EM algorithm was introduced for general latent variable models by Dempster, Laird, & Rubin (1977). Redner and Walker (1984) provides an excellent review of maximum likelihood estimation for finite mixture models, whereas the monograph of McLachlan and Peel (2000) gives full details for a wide range of finite mixture models.

Despite having the advantage of conceptual and computational simplicity, application of the EM approach is difficult when estimating complex models, especially when mathematical derivations are intractable. Several practical problems arise in the likelihood-based approach to mixture model parameter estimation (Frühwirth-Schnatter, 2006). First, it may be difficult to find the global maximum of the likelihood numerically. Second, the likelihood function of mixture models is unbounded and can have many spurious local modes (Kiefer & Wolfowitz, 1956). Similarly, the sample size has to be very large to apply the asymptotic theory of maximum likelihood for mixture models (McLachlan & Peel, 2000). Bayesian techniques, such as Markov chain Monte Carlo (MCMC), which are described next, have become an alternative to address such problems.

2.4.2 Bayesian Approach

Bayesian statistics attempts to formalize and quantify researchers' prior assumptions concerning their research questions. Its main components include a prior distribution, posterior distribution, and likelihood distribution (or function). In the formulation of Bayesian inference, y denotes the observed data, θ denotes model parameters and missing data, and $P(\theta|y)$ denotes probability statements conditioned on observed data. The foundation of Bayesian inference is to set up a *joint probability distribution* $P(\theta,y)$ over all random quantities (Wilks, Richardson, & Spiegelhalter, 1996). For this purpose, we begin with a model providing a joint probability distribution $P(\theta,y)$, which can be expressed as a product of a *prior distribution* $p(\theta)$ and the *likelihood distribution*, $p(y|\theta)$ as follows.

$$P(\theta, y) \propto p(\theta)\, x\, p(y \mid \theta) \qquad (2.3)$$

The prior distribution usually incorporates expert opinions or prior knowledge. Additionally, prior distribution parameters are called *hyper-parameters* when they are not fixed at specific numeric values.

When observed data, y, are available, Bayes theorem is used to determine the distribution of θ conditional on y:

$$P(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta) x\, p(y \mid \theta)}{\int p(\theta)\, x\, p(y \mid \theta) d\theta}$$
$$\propto p(\theta) x\, p(y \mid \theta) \qquad (2.4)$$

This is called the posterior distribution of θ, which is the focus of Bayesian inference. The posterior distributions in simple terms represent the relationship between observed data and prior assumptions (Gill, 2002). A researcher typically uses a likelihood function to quantify purported knowledge concerning the observed data, whereas a probability distribution is placed on prior assumptions to quantify them.

Conjugacy occurs when the posterior distribution follows the same parametric form as the prior distribution. In such cases, inferences may be drawn about one or only a few parameters at a time, and only the marginal posterior distribution needs to be computed for specific parameters of interest. The marginal posterior distribution is obtained by first deriving the joint posterior distribution of all unknowns and then integrating over the unknowns that are not of interest.

2.4.2 Markov chain Monte Carlo Algorithm

Bayesian modeling has appealed to many researchers and practitioners in recent years as a result of development of fast computers along with the availability of powerful computational tools such as Markov chain Monte Carlo (MCMC) algorithms. As mentioned in the previous section, in some situations MCMC is the only means of estimating a model's parameters when required integrals and/or derivatives may not have closed form solutions. MCMC-based estimation offers several other benefits. First, obtaining ability estimates for examinees who answer all questions right or all questions wrong is no longer a problem as it would be with the EM algorithm in the frequentist paradigm. Finite posterior ability estimates can be computed with the help of appropriate priors. Also, likelihoods that are not statistically identifiable can be combined with priors to produce unique posterior distributions (Johnson & Albert, 1999). Additionally, MCMC estimation can also handle several likelihoods in a single analysis. For example, Patz & Junker (1999a) incorporated both dichotomous and polytomous items in the same analysis.

The difficulty of incorporating uncertainty into item parameter estimates has long been a concern with frequentist methods such as the E-M algorithm. Patz and Junker (1999a) noted how the Bayesian approach outlined by Tsutakawa and Soltys (1988) incorporated parameter estimation uncertainty into the standard errors of the estimate. In the context of large scale assessment, particularly in low stakes tests, the matrix of response patterns becomes increasingly sparse as test length increases. Bayesian methods can handle missing data relatively easily within the parameter estimation scheme (Maier, 2002; Patz & Junker, 1999a, 1999b).

The MCMC approach (e.g., Gelfand & Smith, 1990; Gilks, Richardson, & Spiegelhalter, 1996) simulates random samples from the multivariate posterior distribution so that features of the theoretical distribution can be estimated by corresponding features of the resultant random sample (Patz & Junker, 1999b). Based on the original work of Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller (1953), MCMC methods were generalized by Hastings (1970), as the Metropolis-Hastings algorithm. MCMC estimation methods can be thought of as Monte Carlo integration using Markov Chains (Gilks et al., 1996). The key is to create a Markov process whose stationary distribution is the specified *target posterior distribution* $P(\theta|y)$ and run the simulation long enough that the distribution of the current draws achieves stationary. The posterior distribution is summarized by computing statistics based on these draws.

Using previous approaches (Hastings, 1970; Metropolis et al., 1953), Geman and Geman (1984) employed a version of MCMC called Gibbs sampling in physics. Gelfand and Smith (1990) and Gelfand et al. (1990) later introduced this technique to the statistical community as a tool for fitting statistical models.

The general procedure for sampling from the $P(\theta|y)$ is as follows.

- Using a starting point, run independent parallel sequence of an iterative simulation, such as Gibbs sampler or the Metropolis-Hastings algorithm.
- Run the iterative simulation until it reaches the convergence.
- Discard the beginning of the sequence (also known as burn-in period) to eliminate draws that were taken before convergence was achieved.

- Finally, summarize inference about the posterior distribution by treating the set of all iterates from the simulated sequences after burn-in as an identically distributed sample from the target distribution.

Congdon (2005) provides a lucid explanation of the Gibbs sampler both theoretically and mathematically. In order to express the Gibbs sampler in simple terms, let us consider $\theta = (\theta_1, ..., \theta_d)$ with the corresponding univariate conditional distributions of $f_1, \ldots f_d$. The distributions $f_1, \ldots f_d$ are called the full conditional distributions. Similarly, suppose that we can simulate from these full conditional distributions.

$$\theta_i \mid \theta_1, ..., \theta_{i-1}, \theta_{i+1}, ... \theta_d \sim f_i(\theta_i \mid \theta_1, ..., \theta_{i-1}, \theta_{i+1}, ... \theta_d) \qquad (2.6)$$

where $i=1, ..., d$. An iteration of the Gibbs sampler consists of d updates of vectors in iteration t, where each update adjusts one component of θ conditioning on the other $(d-1)$ components. At each iteration t, an ordering of the d subvectors of θ is chosen and, in turn, each θ_i^t is sampled from the conditional distribution given all the other components of θ. Thus each subvector θ_i is updated conditional on the latest values of θ for the other components, which are the iteration t values for the components already updated and the iteration $t-1$ values for the others.

The Gibbs Sampler Algorithm can be summarized by the following expression (Congdon, 2005).

$$1.\ \theta_1^{(t+1)} \sim f_1(\theta_1 \mid \theta_2^{(t)}, \ldots\ldots, \theta_d^{(t)})$$
$$2.\ \theta_2^{(t+1)} \sim f_2(\theta_2 \mid \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots\ldots, \theta_d^{(t)})$$
$$3.\ \theta_3^{(t+1)} \sim f_3(\theta_3 \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_4^{(t)}, \ldots\ldots, \theta_d^{(t)})$$
$$\ldots\ldots$$
$$d.\ \theta_d^{(t+1)} \sim f_d(\theta_d \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots\ldots\ldots, \theta_{d-1}^{(t+1)})$$

Casella and George (1992) defined Gibbs sampling as a "technique for generating random variables from a ...distribution directly, without having to calculate the density" (p. 167). A more technical overview of the Gibbs sampler is provided in Casella and George (1992).

2.4.2.3 Bayesian Estimation in IRT and Mixture IRT Models

Baker and Kim (2004) provided a thorough review of prior research on Bayesian estimation for different parameterizations of item response theory models. In the IRT framework, Gibbs sampling was first used by Albert (1992), which approach was extended for use with numerous IRT models such as those that analyze testlets (Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000), multilevel IRT model (Fox & Glas, 2001), multidimensional models (Béguin & Glas, 2001), and the graded response model (Johnson, 1997). As reflected by the wide use of the Gibbs sampler, Bayesian parameter estimation offers an attractive methodology for experimentation with new and potentially complex IRT models (Kim & Bolt, 2007). In the context of the mixture IRT model, a common MCMC strategy is to sample a class membership parameter for each

examinee along with a continuous latent ability parameter. Practical implementation of MCMC and the Gibbs sample exist in the WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) software package which is described next.

2.4.2.4 WinBUGS and Sampling Methods in WinBUGS

As noted above, WinBUGS (Lunn et al., 2000) is a software package that uses Markov chain Monte Carlo (MCMC) methods to facilitate Bayesian analysis for a wide variety of applications. The material presented in this section draws significantly from Lunn et al. (2000), Spiegelhalter et al. (2003) and Cowles (2004). One of the most attractive features of the WinBUGS is that it is relatively easy to use because it automatically implements numerical sampling and has customized output analysis features that are adequate for most MCMC analytical purposes. On the other hand, as general-purpose software, WinBUGS is not optimized for specific models. Therefore, the time required for completing computations can be rather long and increases for larger datasets and more complex models. For this reason it is generally not practical to use WinBUGS in large scale assessment.

WinBUGS employs different sampling for different types of models. Generally, in simple cases, a conjugate prior distribution is used with a standard likelihood to yield a posterior distribution from which the parameters in the model can be directly sampled. When it is not possible to get the samples directly from the posterior distribution, as in more complex models, some forms of Gibbs sampling and Metropolis-Hastings sampling are used to sample from the posterior distribution.

As mentioned earlier, Gibbs sampling algorithms are used to construct the transition kernels for its Markov chain samplers. Every iteration of a Gibbs sampler involves drawing a new value for each parameter from its full conditional distribution. WinBUGS chooses a method to draw samples during the compilation phase. Consequently, different forms of sampling will be implemented for different parameter types. More information about WinBUGS and its sampling methods can be found in the WinBUGS user manual (Spiegelhalter et al., 2003). It should be noted that the selection of these methods occurs through a process internal to WinBUGS, and is not required to be specified by the user.

2.5 Research Questions

As mentioned in the first chapter, a primary goal of this dissertation was to explore the existence and impact of limitation of commonly used IRT models, particularly their inability to account for the test-taking heterogeneity that might exists in the testing population. Furthermore, inaccuracies in psychometric modeling could have an additional impact on the accurate implementation of educational policies which in turn have consequences for schools, students, parents, and society in general. Therefore, to evaluate the precision of MixIRT models in accurately estimating sample heterogeneity, this study examined different examinee test-taking behaviors using both simulation and empirical analysis. Specifically, this dissertation investigated the following research questions.

- How accurate are the parameter estimates for mixture IRT model when the number of items, the number of examinees, and the proportion of unobserved

test-taking heterogeneity (as represented by % of examinees guessing) are varied?

- How comparable is the precision of parameter estimation in the mixture IRT model to the estimation from the two-parameter logistic (2PL) model and the three-parameter logistic (3PL) model?
- How accurately does Bayesian mixture modeling identify a cluster of examinees who are likely to be *guessers* in a large-scale assessment?
- What is the impact of excluding item responses identified by mixture IRT modeling as coming from *guessers* in proficiency level classification?

The next chapter provides an overview of the methodology and the parameterization of models used in this study which evaluate how varying degrees of test-taking heterogeneity influence the parameter estimation. A subsequent section describes the simulation study design, simulation of data, data analysis and criteria used in parameter recovery. Finally, the mixture IRT model analysis is presented with a real data example.

CHAPTER 3

METHODOLOGY

The primary goal of this research was to evaluate the precision of mixture IRT (MixIRT) models in accurately estimating the differential performance of distinct groups in a sample (i.e., sample heterogeneity). In this study, a MixIRT model was used to investigate different examinee test-taking behaviors through a simulation study that varied (a) sample size, (b) test length, and (c) proportion of guessing examinees. These simulation factors are representative of realistic testing situations and allow an evaluation of how varying degrees of test-taking heterogeneity influence parameter estimation.

The remainder of the chapter is divided into the several sections. The first section provides an overview of the methodology and the parameterization of models used in the simulation study. A subsequent section describes the simulation study design, simulation of data, data analysis and criteria used in a parameter recovery study (i.e., bias, root mean squared error, correlation). The final section discusses a real data example.

3.1 Models

The MixIRT model is the basic theoretical framework of this study. In general, MixIRT models stipulate that different parameter values may apply for different latent classes of examinees in a population. In effect, different IRT models hold true for each of the latent classes. In this thesis, the MixIRT model is formulated from two latent classes where one latent class belongs to the responses following the specified IRT model and

another latent class comprises item responses from the examinees engaged in guessing behaviors.

This study used two types of MixIRT models to characterize various guessing strategies encountered in a typical large scale assessment. In both models, one latent class was the set of responses which followed a two-parameter logistic model. The second latent class represented item responses from examinees using guessing strategies. Two MixIRT models differed on how they incorporated the guessing behaviors. The first model used was the MixIRT model with completely random guessing behavior, labeled hereafter as MixIRT-R. The second model was a MixIRT model comprised of responses from examinees engaged in ability-based guessing behavior, labeled hereafter as MixIRT-A. It should be noted that unless otherwise specified, the label MixIRT model is used to refer to both of these models. What follows is a description of the two MixIRT models used.

3.1.1 Model 1: Mixture IRT model with completely random guessing behavior (MixIRT-R)

The MixIRT model with completely random guessing behavior (MixIRT-R) basically represents a stochastic statistical mixture. Its purpose is to identify probabilistically the random responders from *legitimate* 2-PL responders. The development of MixIRT-R model is related to the idea of the HYBRID model (Yamamoto, 1989) and the effort moderated IRT model (Wise & DeMars, 2006), where the classification of examinee into *guessers* and *non-guessers* is based on probability estimation from the examinee item responses. In addition to some parameterization differences in the model presented below with earlier studies, this study utilized a

different estimation method, i.e. Bayesian estimation. The mathematical representation of MixIRT-R model includes two classes, one of which belongs to a 2-PL model and another belongs to a random guessing class. Accordingly, the probability of answering an item *j* correctly by an examinee *i* is given by:

$$P(X_{ij} \mid g_i, a_j, b_j, \theta_i) = (1-g_i) * \frac{1}{nALT} + g_i \left(\frac{\exp[a_j(\theta_i - b_j)]}{1+\exp[a_j(\theta_i - b_j)]} \right) \quad (3.1)$$

where, nALT represents the number of options in a multiple choice item and g_i represents the group membership (e.g., *guesser* or *non-guesser*). The symbols a_j, b_j and θ_i represent item discrimination, item difficulty, and ability parameters and have the same interpretation as the 3PL parameters of Equation 1.1 in Chapter One. Also, g_i = (0,1) with $g = 0$ as random guessing group (*guessers*) and $g = 1$ as 2PL (*non-guessers*). These parameters can be estimated using the WinBUGS software programs as explained in the next section.

3.1.2 Specification of the MixIRT-R Model in WinBUGS

The parameterization of MixIRT-R model specified in Equation 3.1 is similar to that of the IRT model described in Chapter One. However, there are two additional parameters to be estimated. The first one is g_i which corresponds to the categorical representation of group identification. The hyperparameter of this distribution is parameterized by a Dirichlet distribution which is a conjugate prior for estimating the

proportion of examinees that are likely to be *guessers* and *non-guessers*. Refer to the Appendix A.1 for the WinBUGS code used to implement this model.

3.1.2.1 Specification of model parameter priors

Specification of a prior distribution is one of the most important methodological as well as practical problems in Bayesian inference. Often researchers want "the data to dominate" when there is no prior information and thus attempt to use vague prior distributions (Lambert, Sutton, Burton, Abrams, & Jones, 2005). The *vague prior* is a term used in Bayesian statistics to refer to a prior when the analyst does not have any information about the value of the unknown parameter.

In a simulation study of the impact of vague prior distributions in MCMC using WinBUGS in an IRT model, Lambert et al. (2005) found fewer problems with location parameters (i.e., item difficulty parameters), but found major problems with scale parameters (i.e., item discrimination parameters). This study will use informative priors on the discrimination parameters (a_j) since the existence of the joint posterior distribution is not guaranteed when an improper prior is used (Albert & Ghosh, 2000; Bazan, Branco, & Bolfarine, 2006). It should be noted that this limitation should not be attributed to the Bayesian estimation methods so much as to the modeling flexibility of WinBUGS and obtaining the convergence. However, a vague prior is used for item difficulty parameter as used in earlier studies (e.g., Albert & Ghosh, 2000; Sahu, 2002).

Ability parameters (θ) and item difficulty parameters (b) were each assigned a two-stage normal prior as:

$$\theta_i \sim N(0, \tau_\theta), \quad i = 1, ..., n,$$
$$b_j \sim N(0, \tau_b), \quad j = 1, ..., J. \qquad (3.2)$$

where both τ_θ and τ_b follow the conjugate inverse gamma prior,

$$\tau_\theta \sim IG(\alpha_\theta, \beta_\theta),$$
$$\tau_b \sim IG(\alpha_b, \beta_b). \qquad (3.3)$$

where $\alpha_\theta, \beta_\theta, \alpha_b,$ and β_b are hyperparameters.

These distributional specifications are chosen based on earlier studies in IRT parameter estimation (e.g.,Cao & Stokes, 2008; Hambleton & Swaminathan, 1985; Kim & Cohen, 1998; Patz & Junker, 1999a, 1999b; Sahu, 2002)

3.1.3 Model 2: Mixture IRT model with ability-based guessing (MixIRT-A)

IRT models that handle ability-based guessing are getting increased interest recently (e.g., Cao & Stokes, 2008; Martin, Pino, & De Boeck, 2006). The motivation for this modeling approach is largely based on the assumption that the success of guessing is related to ability. The MixIRT model with ability-based guessing (MixIRT-A) basically represents another variant of a stochastic statistical mixture. The fundamental assumption of this model is that both *guessers* and *non-guessers* take the test, but interact differently with items that are harder for the examinee. Specifically, some examinees utilize their full potential (thoughtful response) to only the relatively easy items and tend to guess answers to test items that are difficult for them.

Based on the theory of test-taking motivation, such as one illustrated by Wise and DeMars (2005), the MixIRT-A model assumes that the amount of effort put forth by an examinee decreases as the task becomes increasingly difficult. In other words, a *guesser* is more likely to guess on items that are difficult, where that difficulty threshold is related to the examinee's own ability parameter. Cao and Stokes (2008) recently used this model to describe partial guessing behaviors, which was labeled as the "..Difficulty-Based Guessing Model". The probability of a correct response for an examinee *i* to test item *j* in this model is given by:

$$p(x_{ij}=1\,|\,\theta_i,a_j,b_j,c_j,\delta_i) = \frac{\exp[a_j(\theta_i-b_j)-\eta_i I(b_j-(\theta_i+\delta_i))\,(a_j(\theta_i-b_j)-c_j)]}{1+\exp[a_j(\theta_i-b_j)-\eta_i I(b_j-(\theta_i+\delta_i))\,(a_j(\theta_i-b_j)-c_j)]} \qquad (3.4)$$

where η_i=1 if examinee *i* is a guesser and 0 otherwise. δ_i is a parameter that measures the difficulty threshold for a guesser to guess. In other words, some examinees may use their full potential or even try illuminating one or two choices before making their guess. Others may not use their full potential, thus guessing on those items that are difficult for them. The indicator function, represented as *I*(...), in Equation 3.4 above becomes 1 only if the difficulty parameter of item *j* is larger than the ability of examinee *i* with some degree of adjustments controlled by the threshold parameter δ_i. The current study allowed δ_i varying among examinees because different examinees have a different threshold in terms of their tendency to guess.

The priors for $\theta_i, a_j, b_j, \tau_\theta, \tau_b$ are the same as those used in the Model 1 above. Appendix A.2 provides the WinBUGS code used to implement MixIRT-A model. The prior for η_i is similar to those used for g_i in MixIRT-R, which corresponds to categorical representation of group identification. The hyperparameter of this distribution is parameterized by a Dirichlet distribution which is a conjugate prior for estimating the probability that a particular examinee is likely to be a *guesser*.

3.2 Simulation Study

Typically a simulation study is used to evaluate the performance of a particular model or method in precisely estimating the model parameters. Accurate estimation of item parameters is important in any psychometric applications such as test equating, item banking, etc. The overarching goal of this study was to evaluate the performance of the MixIRT model in precisely estimating sample heterogeneity, and to study how different testing characteristics influence the estimation of model parameters. Therefore, a simulation study was most appropriate to address these goals. In addition, a simulation study allowed exploration of the impact of guessing behavior on parameter estimation.

Hence, in order to evaluate the extent to which the MixIRT model can precisely recover the item parameters using Bayesian estimation, a parameter recovery study was conducted. The precision of parameter estimation was evaluated in terms of bias, RMSE, and correlation between estimated and simulated parameters. As mentioned earlier, the proposed method provides better item parameter recovery when it produces small bias, small RMSE, and high correlation between estimated and simulated parameters.

3.2.1 Simulation Factors or Study Design

As mentioned earlier, a simulation study allows the evaluation of how different testing characteristics influence the estimation of mixture model parameters. In the context of this study, it is possible to explore the impact of unobserved test-taking heterogeneity (guessing proportion) on parameter estimation. Typical test characteristics, which are encountered in applied testing situations, include sample size, test length, and proportion of guessing. Taking this into account, this study used the factors listed in the Table 3.1, which are commonly used in parameter recovery studies (e.g., Goldman & Raju, 1986; Hulin et al., 1982; Kim & Cohen, 1998).

Table 3.1 Summary of Parameter Recovery Study Factors

Factors	Levels
Sample Size	500, 2000
Test Length	25, 50
proportion of "guessing"	0%, 5%, 10%
Estimation model	MixIRT, 2PL, 3PL

This simulation study used a MixIRT model with simulated random guessing behavior as labeled as MixIRT-R model above. The estimation from 0% guessing serves as a baseline. This study investigated the impact of different guessing proportion (5% and 10%) on parameter estimation. The guessing proportion represents the percentage of examinees who are a *guesser* in a test. The two-parameter logistic (2PL) model was used for generating data. The performance of MixIRT-R and 2PL model was compared with

3PL model because 3PL is commonly used in practice for parameter estimation when guessing behavior is suspected in multiple choice items.

Each condition in this study was replicated 15 times. Although this may appear to be too few replications from a frequentist perspective, this is actually more than the number of replications used in Bayesian IRT-based simulation studies. This reduction in replications is partly a result of the computational intensity of WinBUGS software which can take upto 6 hours to run 25,000 iterations for the item responses with 2000 examinees and 50 items. Examples from the literature have used only five (e.g., Bolt & Lall, 2003) or ten replications (Cao & Stokes, 2008). The general procedures employed to simulate item and ability parameters, and simulation of item responses are presented next.

3.2.2 Generation of Simulated Parameters and Item Responses

The simulation of parameters and item responses was based on typical methods found in IRT literature (Hulin et al., 1982; Kim & Cohen, 1998). Ability parameters were assumed to follow a normal distribution; thus ability parameters were randomly sampled from a standard normal distribution (mean=0, standard deviation=1). Similarly, item discrimination parameters were assumed to follow a lognormal distribution. Thus, discrimination parameters were randomly sampled from a lognormal distribution [$a_i \sim$ lognormal (0,0.3)]. The item difficulty parameters were also assumed to follow a normal distribution. Therefore, difficulty parameters were randomly sampled from a normal distribution with mean of 0 and standard deviation of 0.7. The standard deviation was reduced to slightly less than 1 to avoid too easy or too difficult items.

The a and b parameters were randomly paired with each other. Thus, any nonzero correlations among the item parameters were attributable to chance. These item parameters may be thought of as simulating an idealistic scenario or one that a psychometrician using the 2PL model would hope to obtain.

The probability of a correct response to item j by simulated examinee i was then computed using the two-parameter logistic IRT model (Birnbaum, 1968). A response vector of dichotomous item scores for each examinee was obtained by generating, for each item, a uniform random number (ranging between 0 and 1) and comparing the value with the probability of an examinee of that ability level passing the item. If the computed probability exceeded the random number, then the item score was scored as correct (1); otherwise, it was scored as incorrect (0).

In order to simulate the *guessers*, the item responses from a randomly selected 5% or 10% of total examinees were modified in such a way that their response patterns mimicked guessing behavior. The original data with no guessing (labeled as 0% proportion of guessing) served as baseline data for comparative purposes. The estimation of modified item responses allowed evaluation of the impact of guessing on parameter estimation. Thus, it also showed how 2PL and 3PL models could not account for test-taking heterogeneity.

3.2.3 Parameter Estimation

The item responses simulated or modified above were used as data for item and ability parameter estimation. The primary methodological objective of this study was to compare the estimation from various IRT models (MixIRT, 2PL, 3PL) when model parameters were estimated using computer software WinBUGS. In this program, the

estimations were carried out under Bayesian framework, using sampling procedures such as MCMC and the Gibbs Sampler.

3.2.3.1 Convergence Assessment and Sensitivity Analysis

Evaluating chain convergence is a critical issue in monitoring the simulated states of the Markov chain (Cowles & Carlin, 1996; Kim & Cohen, 1998). In order to view the sampled observations as a sample from the posterior distribution of the model parameters, the sequence of states for the Markov chain should theoretically converge to a stationary distribution. The rate at which this convergence occurs can vary depending on several factors, such as correlations between adjacent states, the sampling algorithm used, and identification problems with the model.

A critical issue for MCMC methods, including Gibbs sampling, is to determine when one can cease sampling and use the results to estimate characteristics of the distributions of parameters of interest (Kim & Cohen, 1998). In this context, the values for the unknown quantities generated by the Gibbs sampler can be graphically and statistically summarized to evaluate for mixing and convergence. Cowles and Carlin (1996) presented a comparative review of convergence diagnostics for MCMC algorithm. The most popular and useful method was that proposed by Gelman and Rubin (1992). This diagnostic measure is implemented in WinBUGS as the Brooks, Gelman, and Rubin (BGR) plot. In this study, five diagnostic measures were used to evaluate the sampler performance: (i) Brooks, Gelman, and Rubin (BGR) diagnostic plots; (ii) Monte Carlo errors; (iii) history plots; (iv) autocorrelation plots; and (v) density plots.

The Gelman-Rubin convergence statistic R compares the ratio of the pooled chain variance to the within chain variance (Gelman & Rubin, 1992). Once convergence is reached, *R* converges to 1. WinBUGS plots 3 items; where the Gelman-Rubin statistic is plotted in red, which is preferred to converge to 1. In blue, the average width of the 80% intervals within each individual chain and the width of the 80% interval of the pooled runs is plotted in green. The blue and green lines should stabilize to some number though it is not necessarily required to be 1.

Monte Carlo error (MC error) is a measure like the standard error of the mean but adjusted for autocorrelation. Generally, autocorrelations for the MCMC sequence that decay slowly as a function of lag imply poor mixing of the MCMC series and could indicate a high-degree of correlations between the parameters or lack of identification of the model. Finally, history and density plots are also useful to monitor the convergence of estimates.

Analysis to evaluate the sensitivity to the initial values and the mixing and convergence of the Gibbs sampler was carried out. The reasonable convergence was reached in each condition by running 3 chains of 25,000 iterations with the first 10,000 discarded as burn-in. For additional replications, however, a single chain of 25,000 iterations was run with the first 10,000 iterations discarded as burn-in period. The estimate of each parameter was based on final 15,000 iterations.

3.2.4 Evaluation Criteria and Analysis of Simulated Data

Three commonly used summary statistics were used as evaluation criteria: bias, Root Mean Squared Error (RMSE), and correlation. Before computing the bias and

RMSE, the estimated parameters were transformed to the same scale as the true parameters. RMSE is the square root of the average of the squared differences between true and estimated parameters across all the items for item parameters and across all the subjects for the ability parameter. For example, in case of item parameter recovery, the RMSE and Bias for each parameter η = a, b are expressed as:

$$RMSE = \sqrt{\sum_{j=1}^{J}\sum_{r=1}^{R}\frac{\left(\hat{\eta}_{jr} - \eta_j\right)^2}{J * R}} \qquad (3.5)$$

$$Bias = \sum_{j=1}^{J}\sum_{r=1}^{R}\frac{\left(\hat{\eta}_{jr} - \eta_j\right)}{J * R} \qquad (3.6)$$

where η_j is the true value and $\hat{\eta}_{jr}$ is the corresponding estimate. J is the total number of items, and R is the number of replications. It should be noted that for ease of interpretation, the results for all J items were combined across the R replications for each simulation condition. Thus, the bias and RMSE presented in the results section are basically the averages of those values across each simulation condition.

Bias index does not indicate in an absolute sense the degree of estimation accuracy. In bias, equal positive and negative errors are cancelled with each other producing a zero bias just as would perfect estimation. The bias then suggests whether there is a systematic tendency to overestimate or underestimate a parameter. A positive bias implies parameter overestimation and a negative bias implies parameter underestimation.

The *correlation* between simulated and estimated parameters was also used as an evaluation criterion because that reflects how well the estimated parameters are correlated with the simulated parameters. The Pearson correlation between estimated and simulated parameter values is given by:

$$r = \frac{\sum_{j=1}^{J}[(\hat{\eta}_j - \bar{\hat{\eta}})(\eta_j - \bar{\eta})]}{\sqrt{\sum_{j=1}^{J}(\hat{\eta}_j - \bar{\hat{\eta}})^2}\sqrt{\sum_{j=1}^{J}(\eta_j - \bar{\eta})^2}} \qquad (3.7)$$

This study also used *classification accuracy* as additional criteria to evaluate how well the MixIRT model classified examinees into a model generated class. For example, to evaluate how well the MixIRT model identified the examinees likely to be in the *guessers* class, the classification accuracy can be expressed in percentage as:

$$Classification\ Accuracy = \frac{Number\ of\ guessers\ identified\ correctly}{Actual\ number\ of\ guessers} X\ 100 \quad (3.8)$$

Since the group membership was modeled as a categorical variable, the *median* was computed for the estimate. The classification accuracy was computed separately for each group (non-guessers and guessers). Because the sample size was different for different groups, weighted classification accuracy was also computed by averaging the classification accuracy values after weighting by sample size.

3.2.5 Simulation Study using Mixture IRT Model with Ability-based Guessing

Although a large part of the simulation study carried out in this dissertation was described in Section 3.2, the assumption in which *guessing* was defined might not be

realistic in all practical testing situations. Thus, the goal of this second simulation study was to use a MixIRT-A model that modeled a different guessing strategy. Specifically, this model accounted for ability-based guessing, as specified as MixIRT-A earlier. Once again, the objective was to show how the simplicity of the 2PL model failed to account for the heterogeneity in testing populations, and to show how Mixture IRT model can account for such heterogeneity. This simulation study, however, simplified the study design by considering only the simulation condition in which the estimation model is varied for a specific test length and sample size. Specifically, the estimation from the 2PL model was compared with the MixIRT-A model for a test of 40 items administered to 1000 examinees. The next chapter provides a summary of simulated item parameters and the results from this analysis.

3.3 Empirical Data Analysis

This study used the data from a large scale assessment obtained from a statewide mathematics assessment administered to Fall 2006 Grade 8 students in a Midwestern state. The data was obtained from over 100,000 students. Although the original test also comprised of some constructed response items, this study used the item responses from 54 multiple-choice items only. Due to the longer computational time required for running MCMC analysis in WinBUGS, samples of 1000 randomly selected test-takers were used. These moderate sized samples were used to carry out empirical analysis. The primary objective of this analysis was to demonstrate an application of the MixIRT model (both MixIRT-R and MixIRT-A) using real data.

*3.3.1 **Analysis of Empirical Data***

The empirical data analysis started with selecting random samples from the statewide assessment mentioned above. First, two samples of size 1000 were selected randomly. Then, WinBUGS was used to estimate model parameters (item and ability) and the group membership of the examinees. In order to demonstrate the application of the MixIRT model in identifying the guessers and showing the impact of guessing on parameter estimation, this study estimated the ability parameters with or without *guessers* in the sample. The calibration was performed twice. First, the model estimated the ability parameters and identified the examinees likely to be from a guesser class. Then, the model was rerun with those guessers removed. It is important to clarify how an examinee was classified as a *guesser* in this study. As noted earlier, the probability of an examinee likely to be a *guesser* was estimated from the item response pattern of the examinee. This probability was actually based on the average over a large number of MCMC iterations. If the probability was equal to or greater than 0.5, the examinee was classified as a *guesser*.

The changes in ability parameter estimation were evaluated in terms of proficiency level classification and the difference between the distribution of ability parameters. The percentage of proficient students is a conceptually simple score-reporting metric that became widely used for school accountability decisions under the *NCLB Act*. In this accountability framework, students are generally classified into four or five different levels based on their performance in a statewide assessment. In most states, there are four proficiency levels: *Advanced*, *Proficient*, *Basic*, and *Below Basic*. This study also used the same convention to represent the proficiency levels. Based on the

ability estimates from a MixIRT model, the distribution of examinees into particular proficiency levels was made as realistic as possible by deriving three cut-scores on the θ-scale that provided the same percentage of examinees into each level reported by this assessment. Evaluation of results from this perspective have potential to provide some policy implications of the findings.

This study used the two independent sample Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1939) to evaluate whether the difference in distribution of θ from the two samples was statistically significant. This nonparametric statistical test is often referred to as distribution free method as it does not rely on assumptions that the data are drawn from a given probability distribution. Specifically, the Kolmogorov-Smirnov test evaluates whether the shapes of the distributions of the two groups are comparable.

In order to test the statistical significance of the differences between proficiency levels classified by two samples, a chi-square test was performed. Pearson's chi-square is the most widely used chi-square test, in which the chi-square statistic is calculated by the difference between each observed and theoretical frequency of each possible outcome. Its formula is given in Equation 3.9 .

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (3.9)$$

where χ^2 is the test statistic that asymptotically approaches a chi-square distribution. O_i is an observed frequency; E_i is an expected frequency under the null hypothesis; n is the number of possible outcomes for each event. Pearson's chi-square statistic is used

to test whether or not an observed frequency distribution differs from a theoretical distribution.

The next chapter provides the results obtained from the simulation study under both guessing models (MixIRT-R and MixIRT-A models described in this chapter). It also outlines the results from the analysis of empirical data from a statewide large scale assessment.

CHAPTER 4

RESULTS

This chapter presents findings from the simulation and real data analyses. Recall that the primary goal of this study was to explore the feasibility of using mixture IRT (MixIRT) models to estimate the differential performance of examinees in different latent classes in a sample i.e., sample heterogeneity. To accomplish this goal, a series of simulation factors were investigated in fully crossed designs, including two sample sizes (500 and 2000 simulees), two test lengths (25 and 50 items), and three proportions of guessing (0%, 5%, 10%). The estimation of model parameters (item and ability) was compared among three models: MixIRT, 2PL, and 3PL.

This chapter is comprised of five sections. The first section summarizes the descriptive statistics of simulated item parameters. The second section presents the convergence of the estimates in WinBUGS because using MCMC sampling to do statistical inference requires convergence of the MCMC chain to its stationary distribution. In the third section, the results obtained from the simulation study under the random guessing model, described as MixIRT-R in Chapter Three, are presented. The fourth section summarizes the results from a simulation study under the ability-based guessing model, described as MixIRT-A in Chapter Three. The final section outlines results from the analysis of empirical data.

4.1 Descriptive Statistics of Simulated Item Parameters

Table 4.1 presents descriptive statistics of the simulated item parameters for both test lengths. Given that these item parameters were randomly selected from specific distributions, the two tests were slightly different in difficulty levels, with the longer test (n=50) being slightly easier than the shorter test (n=25). Since this occurred by a chance due to the difference in test lengths, it should not impact the interpretation of the results.

The discrimination parameters ranged from 0.588 to 1.758 for the shorter test, and from 0.687 to 1.749 for the longer test. The difficulty parameters ranged from -1.896 to 2.086 for the shorter test, and from -2.108 to 2.152 for the longer test. These item parameters are similar to those found in many practical assessments and previous studies. To generalize the results from a simulation study to the practical setting, simulated parameters should be as realistic as possible. Therefore, extreme values of *a*- and *b*-parameters were avoided in the simulation. A complete list of item parameters is listed in Appendix C.1 for test length of 25, and in Appendix C.2 for test length of 50.

Table 4.1 Descriptive Statistics for the Simulated Item Parameters

Test Length	Item Parameter	Mean	Standard Deviation	Maximum	Minimum
25	*a*	1.030	0.313	1.758	0.588
	b	-0.120	0.903	2.086	-1.896
50	*a*	1.076	0.243	1.749	0.687
	b	-0.266	0.797	2.152	-2.108

4.2 Evaluation of Parameter Estimate Convergence

Using MCMC sampling to do statistical inference requires convergence of the MCMC chain to its stationary distribution. Five diagnostic measures, as described in Chapter Three, were used to evaluate convergence: (i) Brooks, Gelman, and Rubin (BGR) diagnostic plots; (ii) Monte Carlo errors; (iii) history plots; (iv) autocorrelation plots; and (v) density plots. It should be noted that no diagnostics can prove convergence, but these multiple criteria provide the indication that convergence might have occurred. These criteria may help in evaluating MCMC convergence to ensure that the samples are fairly representative of the underlying stationary distribution of the Markov chain.

Figure 4.1 presents BGR diagnostic plots, history plots, autocorrelation plots, and density plots for discrimination parameter of a randomly-selected item estimated using the MixIRT-R model. This item has true discrimination parameter of 1.757 and the estimated parameter of 1.818. Similar plots for estimation of difficulty parameter of a randomly selected item and plots for estimation of ability parameters for a randomly selected examinee are given in Appendix B. These plots were chosen from the dataset in which the guessing percentage was 10% for a sample size of 500 and a test length of 25. This condition was chosen here because a small sample size and a short test generally yielded poor parameter recovery and sometimes produced chains that had difficulty in arriving at convergence. Evaluation of convergence from this condition may capture the representative findings from this study.

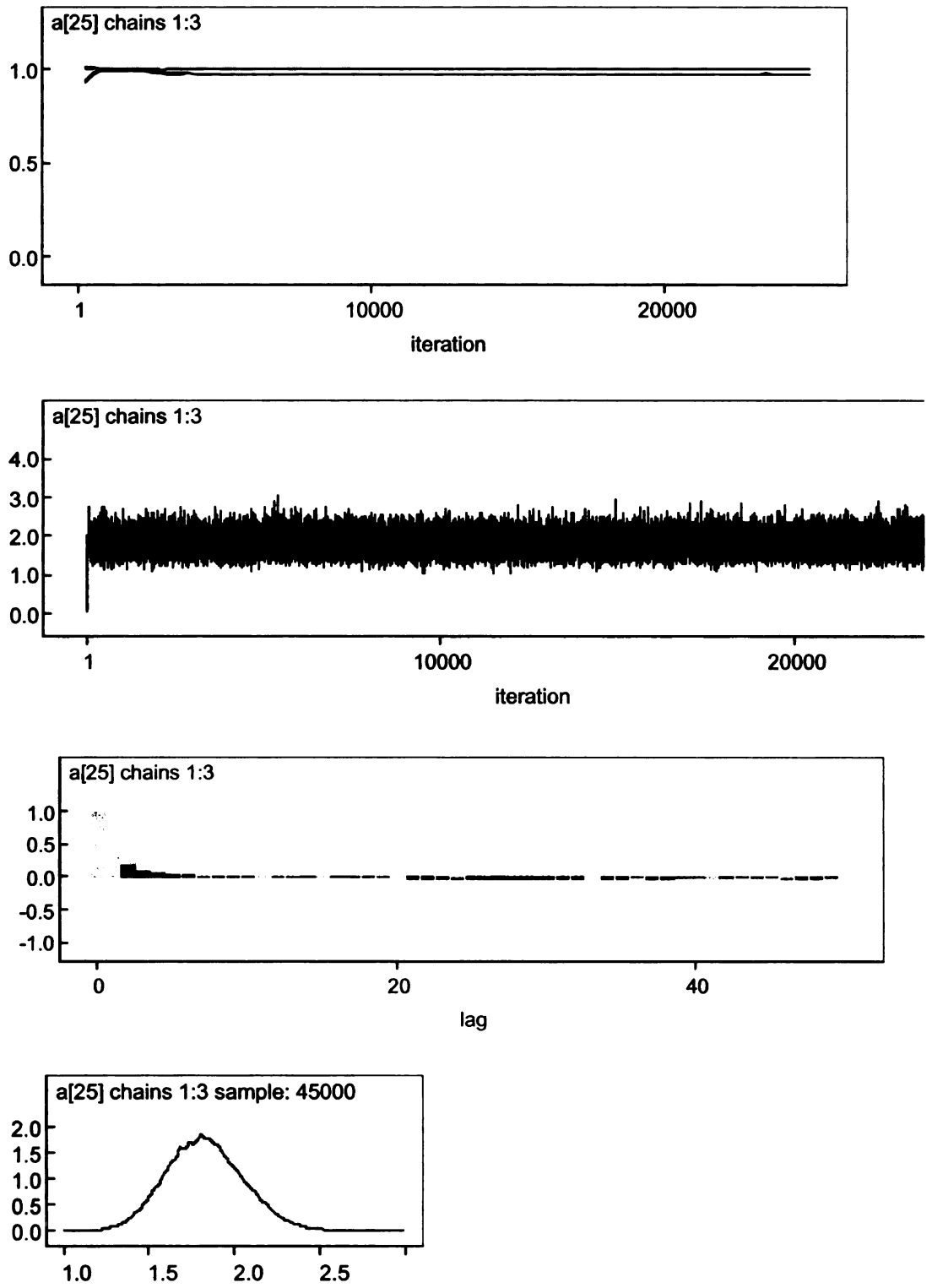


Figure 4.1 Sample plots for convergence assessment of discrimination parameter estimate (From Top: BGR plot, History plot, Autocorrelation plot, Density plot)

BGR Plots

The BGR plot shown in Figure 4.1 indicates that the Gelman-Rubin statistic, which is plotted in red, has converged to 1. The average width of the 80% intervals within each individual chain is plotted in blue and the width of the 80% interval of the pooled runs is plotted in green. Both blue and green lines are stabilized to some number indicating adequate convergence of the chains. It is important to note that the colors shown in BGR plot might be difficult to distinguish in gray scale prints.

Monte Carlo error

Monte Carlo error facilitates the evaluation of convergence by suggesting that how long the simulation should be run to ensure adequate convergence. Table 4.2 presents the descriptive statistics of the MixIRT estimate for randomly selected item responses. For convenience of illustration, only results for the first five items and the first five examinees are shown.

As a rule of thumb, the simulation should be run until the Monte Carlo error for each parameter of interest is less than about 5% of the sample standard deviation (Spiegelhalter et al., 2003). From the Table 4.2, it is clear that the Monte Carlo error is less than $1/20^{th}$ of the standard deviation of the estimate indicating adequate convergence.

History Plots

The history plots in Figure 4.1 suggest convergence has been achieved since three chains essentially overlapped each other and could not be easily differentiated. Furthermore, the convergence seems have been reached well before the burn-in period of 10000 used in this study.

Table 4.2 Descriptive Statistics of MixIRT Estimates for Selected Parameters

Node	Mean	Standard Deviation	MC Error
a_1	0.9335	0.1462	0.0020
a_2	1.0500	0.1536	0.0023
a_3	0.9025	0.1463	0.0018
a_4	1.1600	0.1711	0.0035
a_5	1.5760	0.2484	0.0052
.....			
b_1	0.2472	0.1262	0.0021
b_2	0.1520	0.1164	0.0020
b_3	0.2184	0.1295	0.0021
b_4	-1.1760	0.1808	0.0047
b_5	-1.7730	0.2218	0.0063
.....			
θ_1	1.8550	0.5176	0.0041
θ_2	0.5016	0.3885	0.0035
θ_3	0.8399	0.4132	0.0036
θ_4	-0.5894	0.3922	0.0041
θ_5	0.2236	0.3801	0.0032

Autocorrelation Plots

As shown in Figure 4.1, autocorrelations for the MCMC sequence decayed rapidly as a function of lag, which indicates that there was good mixing of the MCMC series. This shows a lack of correlations between the parameters and indicates satisfactory convergence.

Density Plots

The density plots of Figure 4.1 also suggested the convergence of estimates because the density resembled the appropriate distribution for discrimination parameter.

Thus, after evaluating all the diagnostic measures, adequate convergence was achieved. The additional plots given in the appendix also suggested the adequate convergence.

4.3 Results of MixIRT-R Model Simulation Analyses

4.3.1 Results from the Parameter Recovery Study

Recall that the parameter estimate was evaluated by comparing the estimated model parameters (i.e., discrimination, difficulty, and ability parameters) to the true (simulated) parameters. As mentioned earlier, this study used bias, RMSE, and correlation between estimated and simulated parameters as evaluation criteria. The results are presented both numerically and graphically.

Table 4.3 below summarizes the bias and RMSE of item difficulty parameter *(b)* estimates that were described in Equation 3.6 and Equation 3.5 respectively. Similarly, Table 4.4 summarizes the bias and RMSE of item discrimination parameter *(a)* estimates. The Bias and RMSE values for b and a parameters are also plotted separately for test lengths of 25 and 50. Only selected plots are presented here, and the remaining plots can be found in Appendix D.

Figure 4.2 shows average bias for recovery of item difficulty parameters when test length is 25, whereas Figure 4.3 shows average bias for recovery of item difficulty parameters when test length is 50. The plots corresponding to RMSE values for recovery of item discrimination parameters are shown in Figure 4.4 and Figure 4.5 for test lengths of 25 and 50 respectively. It should be noted that the labels on the x-axis reflect guessing proportion and sample size. For example, 10P500 indicates a sample size of 500 simulees when the percentage of simulees that were guessing was 10%.

It can be noticed in Figure 4.4, that when a 2PL model was used with test length of 25 and sample size of 500, the RMSE increased from 0.129 to 0.152 when the percentage of guessers increased from 0% to 5%. The RMSE increased further to a value of 0.192 when the simulated proportion of examinee guessing increased to 10%. Similarly, the RMSE increased from 0.130 to 0.146 for a 5% guessing percentage and to 0.174 for 10% guessing. Both bias and RMSE values were generally lower with the MixIRT-R model than with the 2PL model. However, both bias and RMSE tended to increase for both models when the percentage of guessers increased to either 5% or 10%.

One of the primary objectives in varying study factors like test lengh and sample size was to evaluate their capacity to recover stipulated item and person parameters. These results show that smaller bias and RMSE were produced by larger sample sizes. The only exception to this sample size finding occurred with the use of the 2PL model for a 50-item test when there were 2000 simulees. For example, in a condition with a 25-item test and 5% guessing percentage, the RMSE value dropped from 0.152 to 0.110 with the 2PL model when sample size was increased from 500 to 2000. Additionally, the RMSE value dropped from 0.141 to 0.080 in MixIRT-R model estimation when sample size was increased from 500 to 2000. No clear pattern of results existed for bias when test length was increased from 25 to 50.

Table 4.3 Bias and RMSE of Item Difficulty Parameter Estimates

IRT Model	Number of Items	Sample Size	0% Guessing Proportion		5% Guessing Proportion		10% Guessing Proportion	
			BIAS	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
			Mean	Mean	Mean	Mean	Mean	Mean
2PL	25	500	0.004	0.129	0.058	0.152	0.096	0.192
	25	2000	-0.002	0.069	0.061	0.110	0.103	0.159
	50	500	0.005	0.127	0.056	0.146	0.088	0.174
	50	2000	0.006	0.062	0.063	0.104	0.245	0.252
MixIRT	25	500	-0.012	0.130	-0.036	0.141	-0.058	0.156
	25	2000	-0.008	0.069	-0.030	0.080	0.007	0.156
	50	500	-0.001	0.128	-0.016	0.134	-0.024	0.137
	50	2000	0.004	0.061	-0.014	0.065	0.019	0.070

Table 4.4 Bias and RMSE of Item Discrimination Parameter Estimates

IRT Model	Number of Items	Sample Size	0% Guessing Proportion		5% Guessing Proportion		10% Guessing Proportion	
			BIAS	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
			Mean	Mean	Mean	Mean	Mean	Mean
2PL	25	500	0.021	0.144	0.045	0.160	0.080	0.202
	25	2000	0.031	0.079	0.065	0.115	0.097	0.170
	50	500	0.032	0.135	0.054	0.163	0.090	0.202
	50	2000	0.038	0.077	0.074	0.116	0.105	0.168
MixIRT	25	500	0.016	0.144	0.015	0.143	0.020	0.151
	25	2000	0.029	0.079	0.035	0.085	0.095	0.178
	50	500	0.031	0.135	0.028	0.141	0.039	0.147
	50	2000	0.037	0.077	0.042	0.081	0.042	0.086

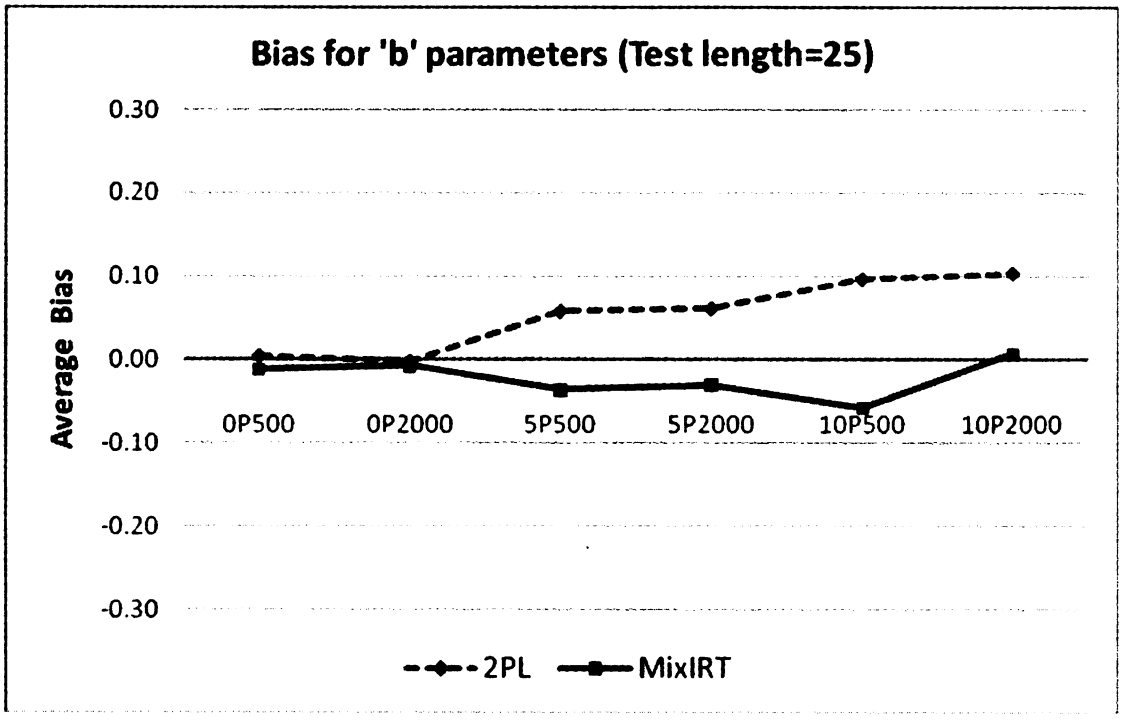


Figure 4.2 25-item test average bias results for difficulty parameter estimates

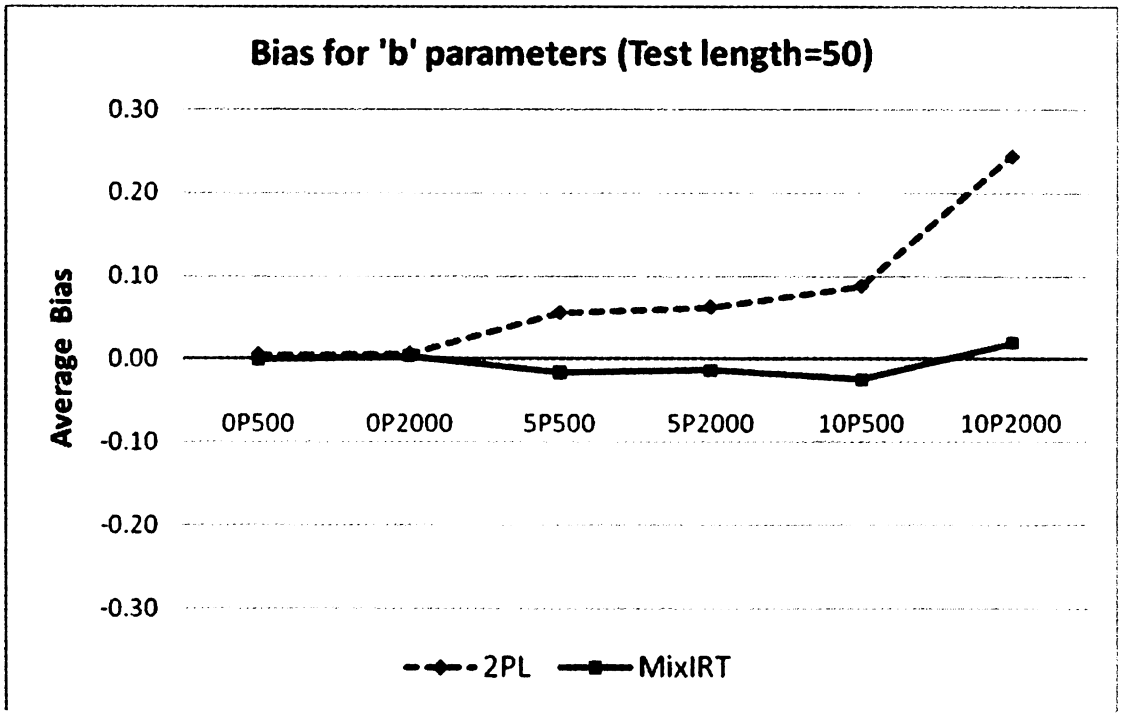


Figure 4.3 50-item test average bias results for difficulty parameter estimates

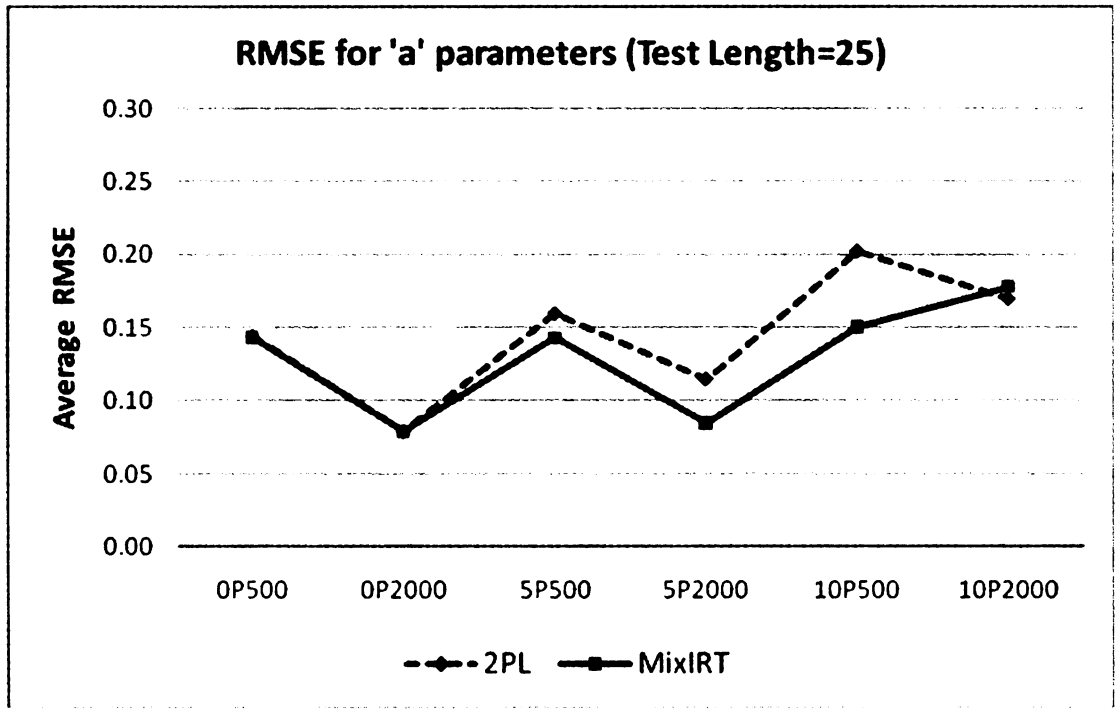


Figure 4.4 25-item test average RMSE results for discrimination parameter estimates

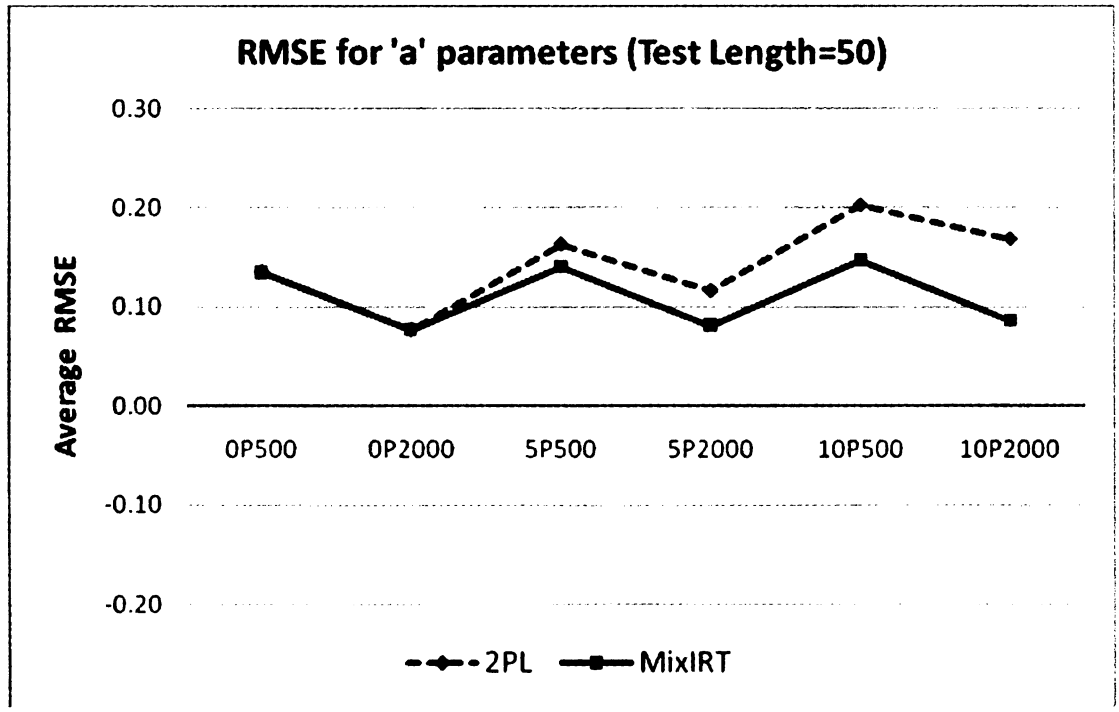


Figure 4.5 50-item test average RMSE results for discrimination parameter estimates

Table 4.5 summarizes the average correlations between true (simulated) and estimated item parameters. These values are presented graphically in Figures 4.6 and 4.7 for discrimination parameters and in Figures 4.8 and 4.9 for difficulty parameters. Clearly, larger correlations were associated with larger sample sizes for both 2PL and MixIRT-R models. The impact of guessing was strong in the recovery of *a* parameters for the 2PL model. For example, correlations between true and estimated *a* parameters dropped from 0.877 to 0.807 when the proportion of guessers increased from 5% to 10% with the 2PL model as shown in Table 4.5. The correlations were similar (about 0.9) for both the 2PL and the MixIRT-R model when no guessers were included in the sample for the condition with the sample size of 500 and test length of 25.

Table 4.5 Correlations between True and Estimated Item Parameters

IRT Model	Number of Items	Sample Size	0% Guessing Proportion		5% Guessing Proportion		10% Guessing Proportion	
			$r_{aa'}$	$r_{bb'}$	$r_{aa'}$	$r_{bb'}$	$r_{aa'}$	$r_{bb'}$
2PL	25	500	0.909	0.989	0.877	0.985	0.807	0.976
	25	2000	0.972	0.997	0.932	0.993	0.839	0.978
	50	500	0.866	0.985	0.779	0.982	0.665	0.974
	50	2000	0.965	0.997	0.907	0.992	0.770	0.979
MixIRT	25	500	0.909	0.989	0.906	0.988	0.902	0.987
	25	2000	0.971	0.997	0.967	0.996	0.956	0.994
	50	500	0.867	0.985	0.852	0.984	0.842	0.984
	50	2000	0.964	0.997	0.961	0.996	0.956	0.996

Note: $r_{aa'}$ *is the correlation between true (a) and estimated (a') parameters*

$r_{bb'}$ *is the correlation between true (b) and estimated (b') parameters*

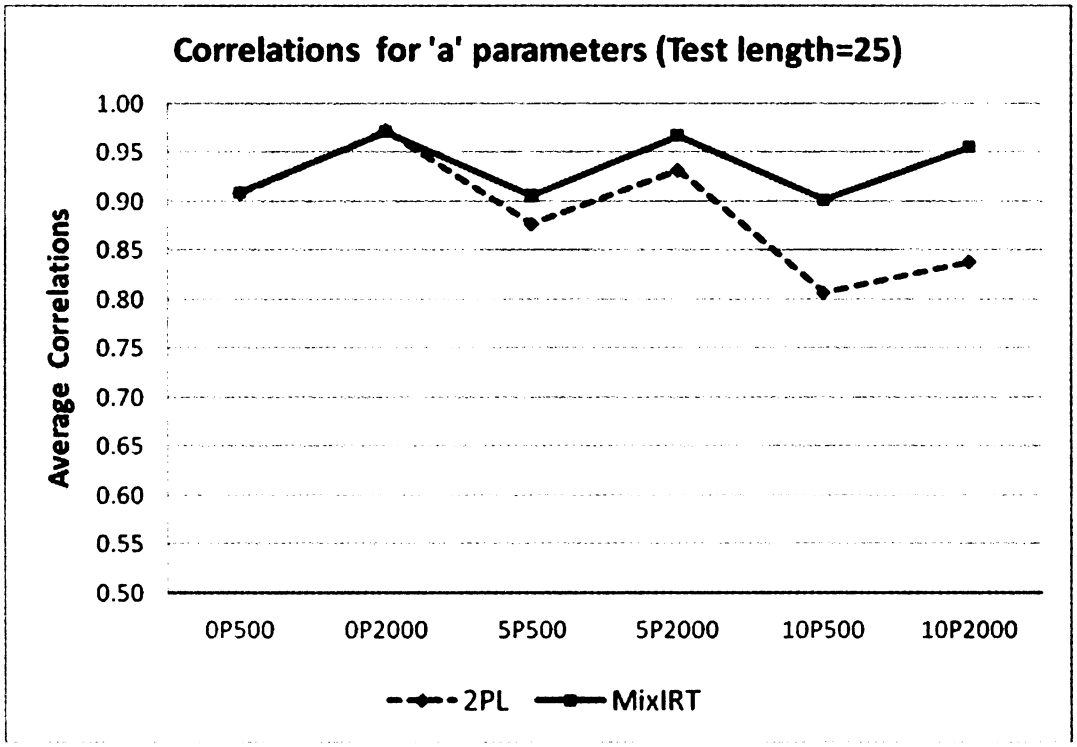


Figure 4.6 25-item test average correlations between true and estimated *a*-parameters

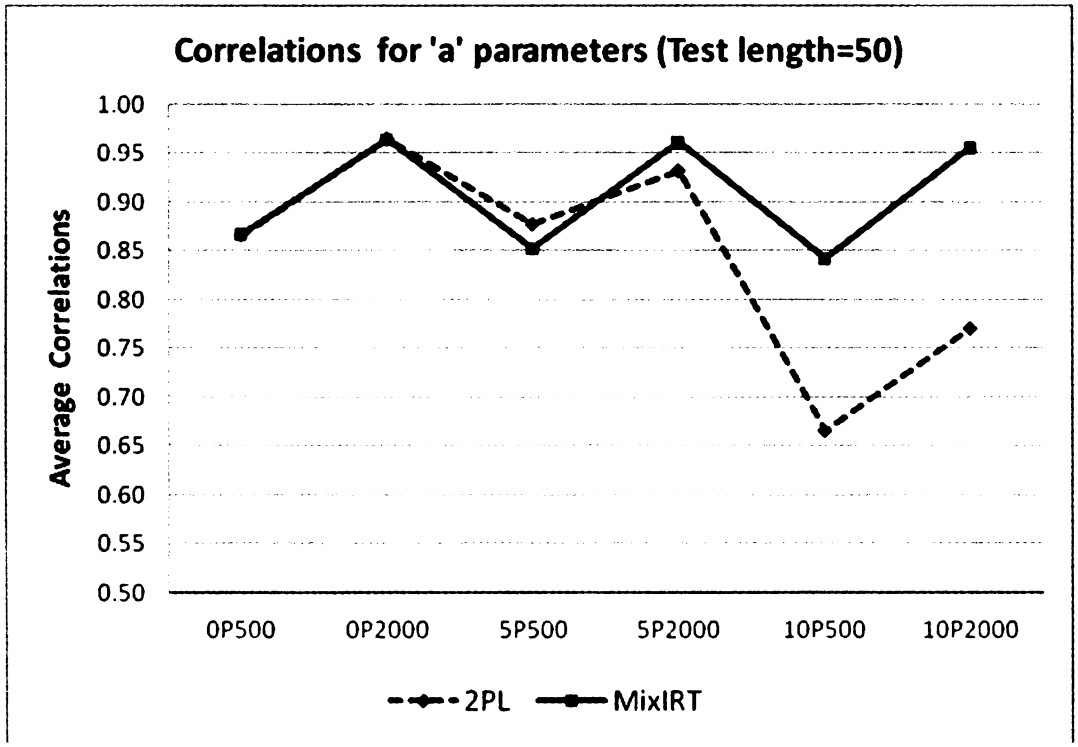


Figure 4.7 50-item test average correlations between true and estimated *a*-parameters

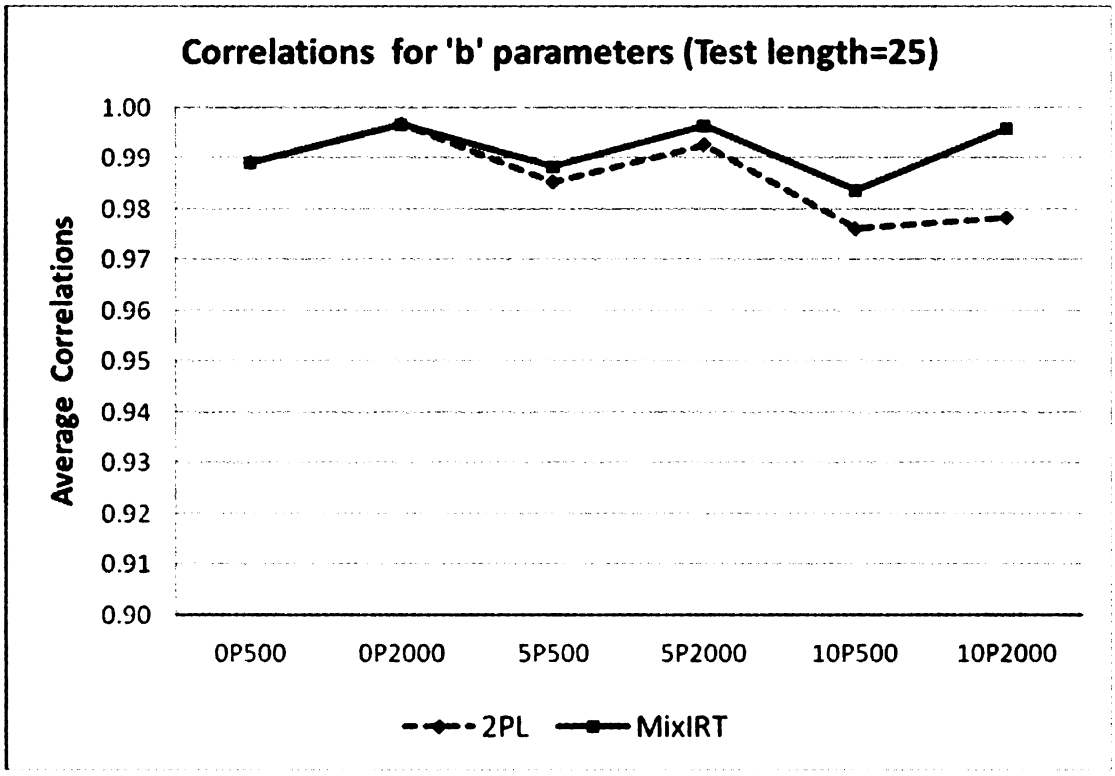


Figure 4.8 25-item test average correlations between true and estimated *b*-parameters

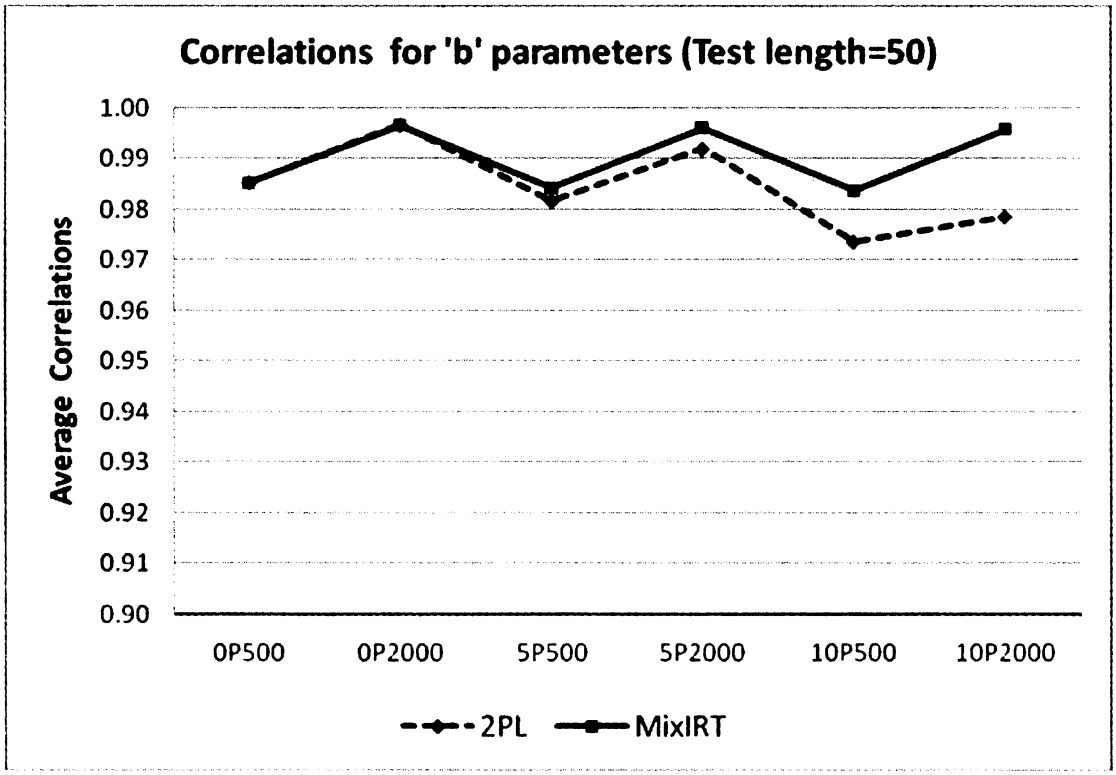


Figure 4.9 50-item test average correlations between true and estimated *b*-parameters

The results pertaining to the recovery of item parameters are also displayed using scatterplots in Figures 4.10 to 4.13. The results presented in these figures include recovery of both 2PL and MixIRT-R models when the percentage of guesser in the sample was 10%. Figures 4.10 and 4.11 are the scatterplots of true and estimated parameters for the conditions of sample size (N) of 500 and test length (n) of 25 for 2PL and MixIRT-R models respectively. Similarly, Figures 4.12 and 4.13 represent the scatterplots for sample size of 2000 and test length of 50 for 2PL and MixIRT-R models respectively.

In a scatterplot, each dot represents the estimated value of a particular parameter for the given value of true parameter. Ideally, for a perfect recovery all dots should fall over the line passing through the origin. Clearly, consistent with the findings presented earlier, the recovery of difficulty parameters (b) was better than that of the discrimination parameters (a) in both models. The recovery of both parameters was better in the MixIRT-R model than in the 2PL model.

The results regarding the recovery of ability (θ) parameters are summarized numerically in Tables 4.6 and 4.7, and graphically in Figures 4.14 and 4.15. Only sample plots are included in this chapter. The results in these tables were similar for both the 2PL and the MixIRT-R models, but more clearly distinct for the 3PL model. The findings indicate that guessing did not have a meaningful impact on correlations between estimated and simulated parameter estimates. Specially, correlations between estimated and simulated θ parameters were at least 0.9.

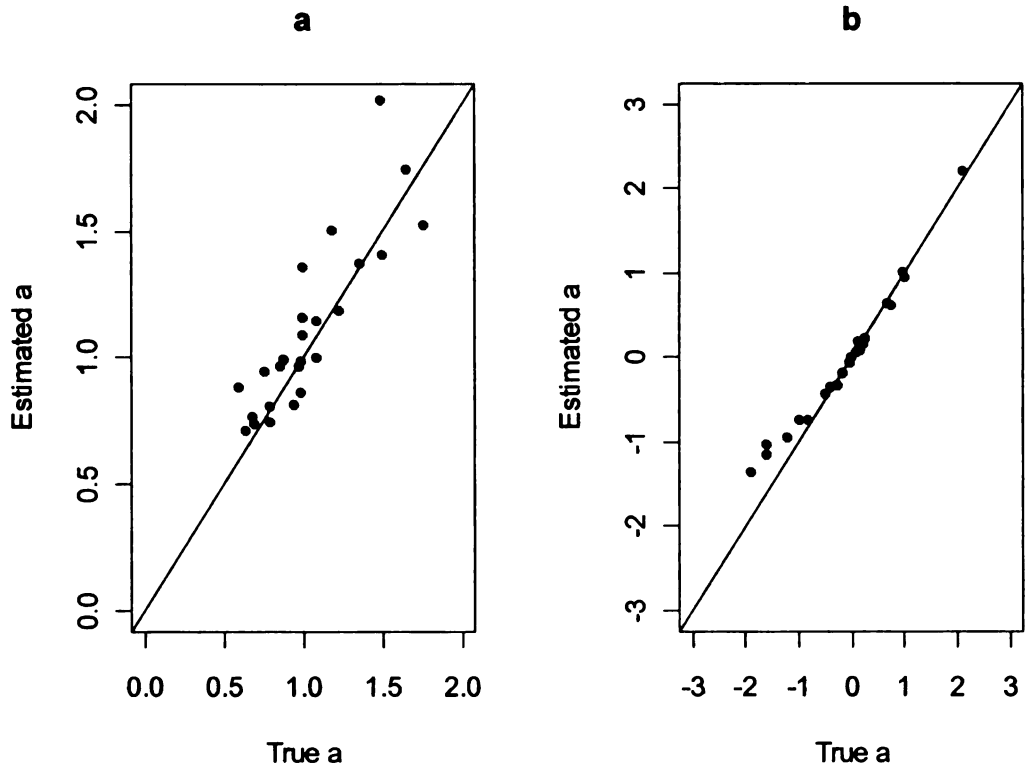


Figure 4.10 Recovery of *a* and *b* parameters in the 2PL model for sample size of 500 and test length of 25 and 10% proportion of guessers

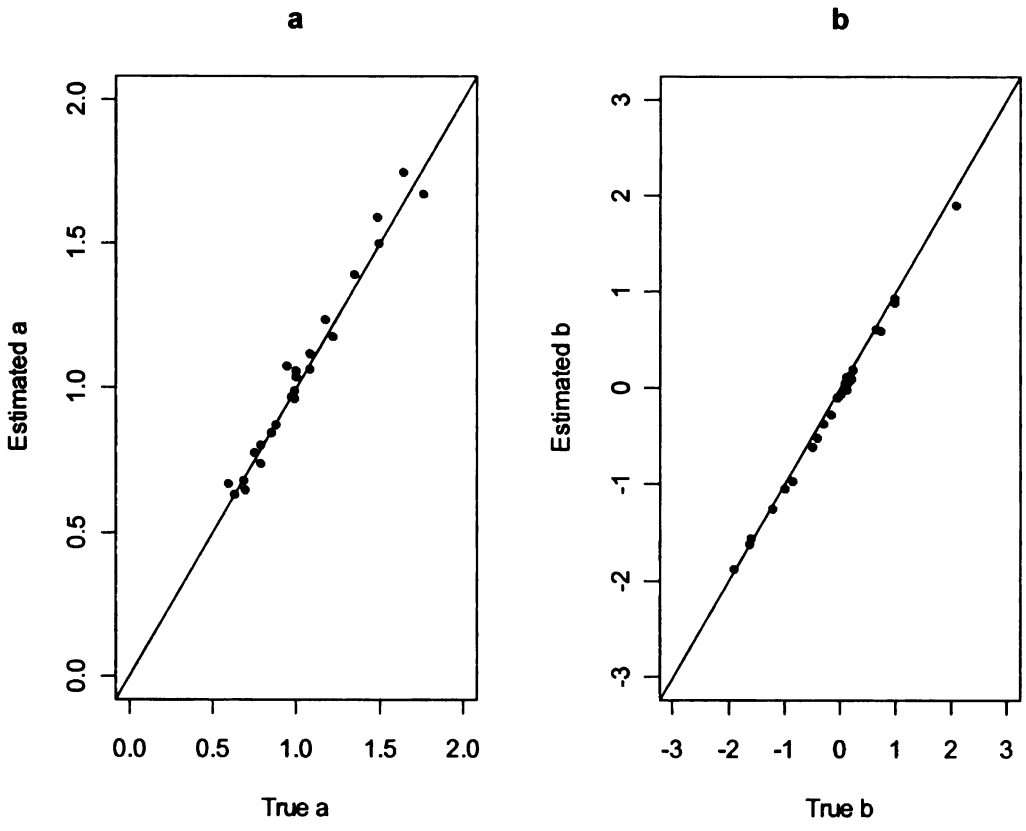


Figure 4.11 Recovery of *a* and *b* parameters in the MixIRT model for sample size of 500 and test length of 25 and 10% proportion of guessers

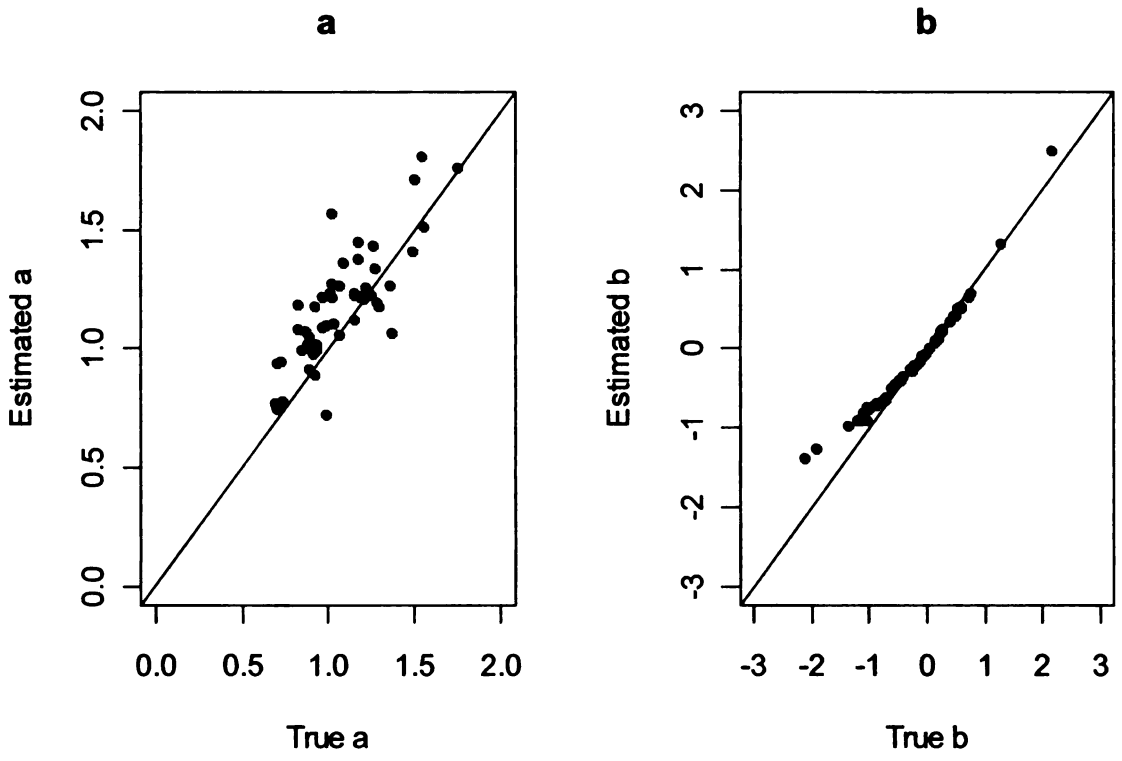


Figure 4.12 Recovery of *a* and *b* parameters in the 2PL model for sample size of 2000 and test length of 50 and 10% proportion of guessers

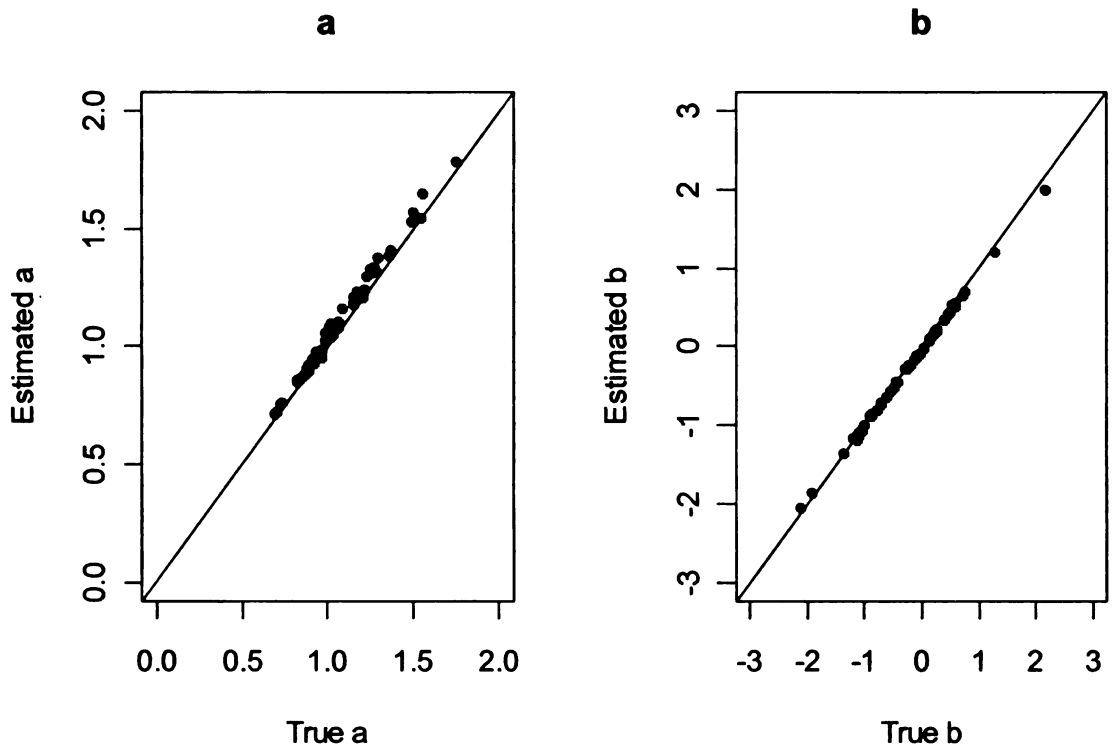


Figure 4.13 Recovery of *a* and *b* parameters in the MixIRT model for sample size of 2000 and test length of 50 and 10% proportion of guessers

Table 4.6 Bias and RMSE of Ability Parameter Estimates for all Simulation Conditions

IRT Model	Number of Items	Sample Size	0% Guessing Proportion		5% Guessing Proportion		10% Guessing Proportion	
			BIAS	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
			Mean	Mean	Mean	Mean	Mean	Mean
2PL	25	500	-0.004	0.402	-0.058	0.404	-0.096	0.414
	25	2000	0.002	0.404	-0.061	0.408	-0.103	0.417
	50	500	-0.005	0.290	-0.056	0.296	-0.088	0.307
	50	2000	-0.006	0.292	-0.063	0.297	0.011	0.304
3PL	25	500	-0.315	0.508	-0.315	0.506	-0.316	0.506
	25	2000	-0.212	0.451	-0.229	0.458	-0.242	0.468
	50	500	-0.294	0.411	-0.297	0.411	-0.295	0.409
	50	2000	-0.214	0.358	-0.222	0.356	0.000	0.307
MixIRT	25	500	0.012	0.404	0.036	0.417	0.058	0.429
	25	2000	0.008	0.404	0.030	0.415	-0.007	0.439
	50	500	0.001	0.291	0.016	0.303	0.024	0.312
	50	2000	-0.004	0.293	0.014	0.306	0.000	0.316

Table 4.7 Correlations between Simulated and Estimated Ability Parameters for all Simulated Conditions

IRT Model	Number of Items	Sample Size	0% Guessing Proportion	5% Guessing Proportion	10% Guessing Proportion
			$r_{\theta\theta'}$	$r_{\theta\theta'}$	$r_{\theta\theta'}$
2PL	25	500	0.910	0.910	0.909
	25	2000	0.913	0.913	0.912
	50	500	0.955	0.954	0.953
	50	2000	0.955	0.955	0.952
3PL	25	500	0.909	0.909	0.908
	25	2000	0.912	0.912	0.911
	50	500	0.953	0.953	0.953
	50	2000	0.954	0.955	0.953
MixIRT	25	500	0.909	0.898	0.889
	25	2000	0.913	0.906	0.902
	50	500	0.954	0.948	0.943
	50	2000	0.955	0.949	0.945

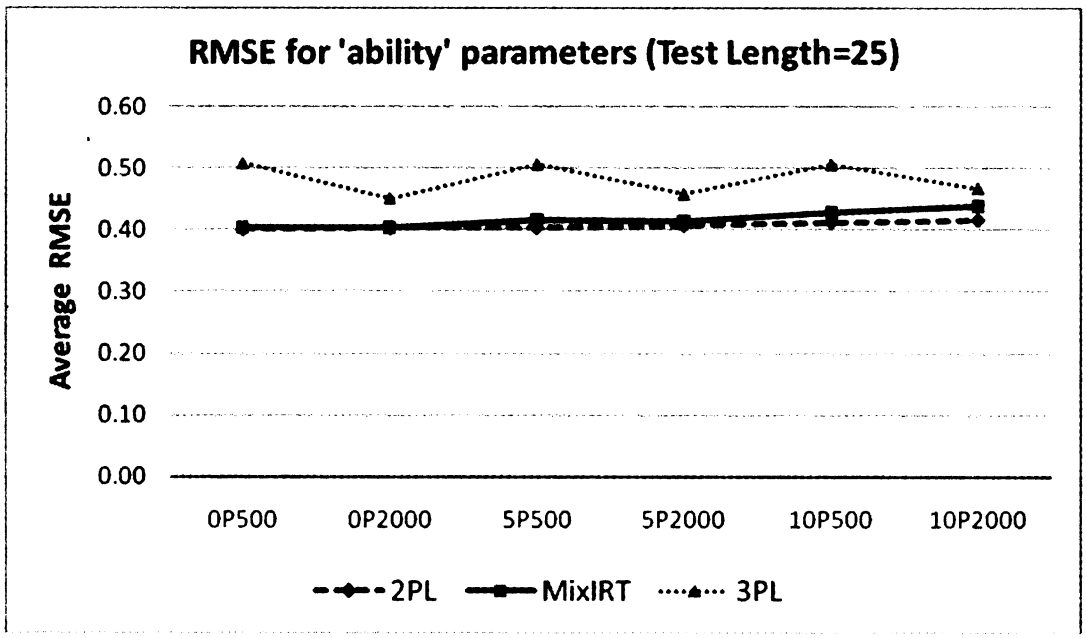


Figure 4.14 25-item test average RMSE results for ability parameter estimates

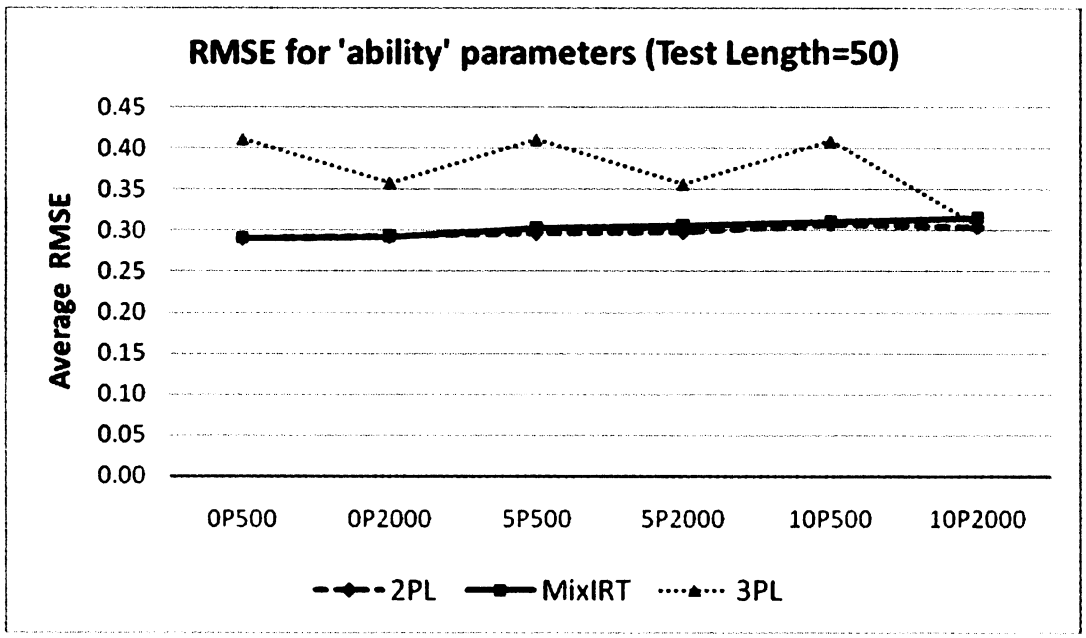


Figure 4.15 50-item test average RMSE results for ability parameter estimates

4.3.2 Classification Accuracy of the MixIRT-R Model

As noted previously, one of the purposes of this study was to investigate the accuracy with which Bayesian estimation of the MixIRT model can correctly identify guessers in a sample. Specifically, the goal was to evaluate the accuracy of classifying examinees into *guesser* and *non-guesser* groups. This research purpose can be addressed only through a simulation study because in real assessment it is impossible to make such a conclusion. Therefore, using estimates from the parameter recovery study described earlier, classification accuracy is ascertained by the extent to which simulees are correctly categorized as guessers or non-guessers.

Table 4.8 provides results of weighted and unweighted classification accuracy for different guessing proportions when using the MixIRT-R model. Overall, the classification accuracy was over 98% for the *non-guessing* group and approximately 90% for the *guessing* group. The weighted classification accuracy, computed by weighting the results by the associated sample size, was 97.20% when sample size was 500 and the guessing proportion was 10%. This accuracy increased to 98.06% when sample size increased from 500 to 2000 simulees. Similarly, when the proportion of guessing was 5%, the weighted classification accuracies were 96.92% and 98.00% for sample size of 500 and 2000 respectively. Interestingly, both classification accuracy and weighted classification accuracy were 100% when no guessers were present (labeled as 0%).

Table 4.8 Classification Accuracy in MixIRT-R Model

Proportion of guessing	Sample size	True Class (Group*)	Average Estimated Guesser %	Classification Accuracy %	Weighted** Classification Accuracy %
10%	500	NG	1.22	98.78	97.20
		G	83.00	83.00	
	2000	NG	1.27	98.73	98.06
		G	85.20	85.20	
5%	500	NG	0.76	99.24	96.92
		G	76.00	76.00	
	2000	NG	0.96	99.04	98.00
		G	78.20	78.20	
0%	500	NG	0	100	100
		G	NA	NA	
	2000	NG	0	100	100
		G	NA	NA	

*NG = Non-guessers, G = Guessers

** Weighted by sample size

4.4 Results from Simulation Analyses using MixIRT-A Model

As noted previously, the guessing factor may not be easy to model in practice, and hence the only way to illustrate it is through a simulation study. The goal of this second simulation study was to use a MixIRT model to incorporate a different guessing strategies (i.e., the assumption of ability-based guessing), that can be modeled using MixIRT-A of Chapter 3. Therefore, the second simulation study showed how the 2PL model is limited in its parameter estimation accuracy because it cannot account for sample heterogeneity. However, this simulation design was simplified by considering

only conditions in which the estimation model was varied for a specific test length and sample size. Specifically, estimation results using the 2PL and MixIRT-A models for 40-item tests administered to 1000 examinees were compared.

Table 4.9 summarizes descriptive statistics of simulated item parameters used in second simulation study. The a parameters ranged from 0.633 to 1.897 with mean of 1.015 and standard deviation of 0.274. The b parameters ranged from -2.274 to 1.945 with mean of 0.093 and standard deviation of 0.855. A complete list of item parameters are listed in Appendix C-3.

Table 4.9 Descriptive Statistics of Simulated Item Parameters in MixIRT-A Model

Item Parameter	Mean	Standard Deviation	Maximum	Minimum
a	1.015	0.274	1.897	0.633
b	0.093	0.855	1.945	-2.274

The same five diagnostic measures used in the first simulation study were used to evaluate the convergence of the estimates. The recovery of item and ability parameters was evaluated using RMSE and correlations between estimated and simulated parameters. The results from this simulation study are presented in Tables 4.10 to 4.14 and Figures 4.16 to 4.18.

Table 4.10. RMSE of Discrimination and Difficulty Parameter Estimates using MixIRT-A Model

IRT Model	No Guessers		Guessers	
	a	*b*	*a*	*b*
	Mean	Mean	Mean	Mean
2PL	0.100	0.096	0.187	0.199
MixIRT	0.102	0.097	0.133	0.084

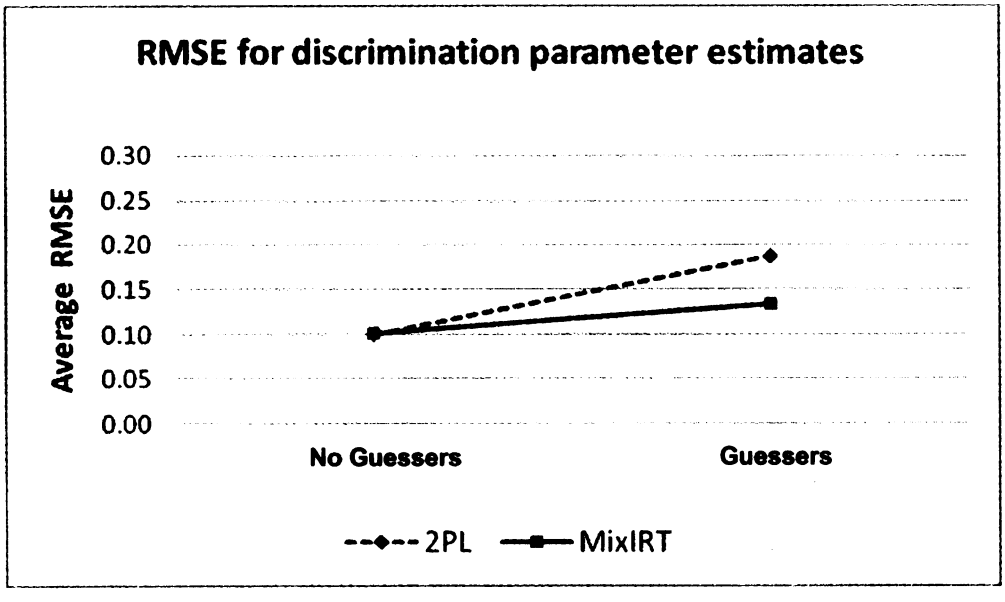


Figure 4.16 RMSE of discrimination parameter estimates in MixIRT-A model

The recovery of discrimination and difficulty parameters indicates that both the 2PL and MixIRT-A models produced comparable results when no guessers were present, i.e. no heterogeneity existed. However, when some simulees were simulated as guessing on items that were likely to be difficult for their given ability level, the MixIRT-A model outperformed the 2PL model. This was reflected by smaller RMSE and larger correlations between estimated and simulated item parameters. Recovery of difficulty

parameters was better than that of discrimination parameters, and guessing had a large impact on the discrimination parameter estimates. For example, in the presence of guessing, the correlation between estimated and simulated discrimination parameter dropped from 0.949 to 0.705 in the 2PL model. However, guessing did not have much impact on the recovery of difficulty parameters. The correlations between true and estimated parameter remained fairly high with values greater than 0.98 in both models.

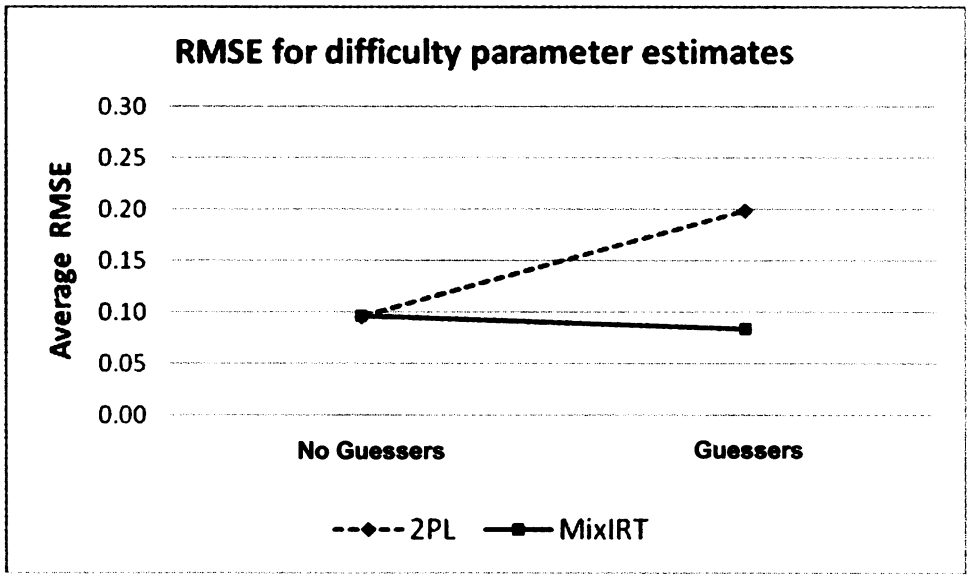


Figure 4.17 RMSE of difficulty parameter estimates in MixIRT-A model

Table 4.11 Correlation of Discrimination and Difficulty Parameter Estimates using MixIRT-A Model

IRT Model	No Guessers		Guessers	
	$r_{aa'}$	$r_{bb'}$	$r_{aa'}$	$r_{bb'}$
2PL	0.949	0.993	0.705	0.988
MixIRT	0.948	0.994	0.886	0.995

The recovery of ability parameters was also evaluated in terms of RMSE and correlations. Table 4.12 and Figure 4.14 show the recovery of ability parameter estimates. In the case of the 2PL model, RMSE increased from 0.325 to 0.411 in presence of guessing. However, the increase in RMSE for the MixIRT-A model was small and increased from 0.326 to 0.343. When guessing was allowed, the correlation decreased from 0.942 to 0.917 in the 2PL model and from 0.942 to 0.929 in the MixIRT-A model.

Table 4.12. RMSE of Ability Parameter Estimates in MixIRT-A Model

IRT Model	No Guessers	Guessers
	Mean RMSE	Mean RMSE
2PL	0.325	0.411
MixIRT	0.326	0.343

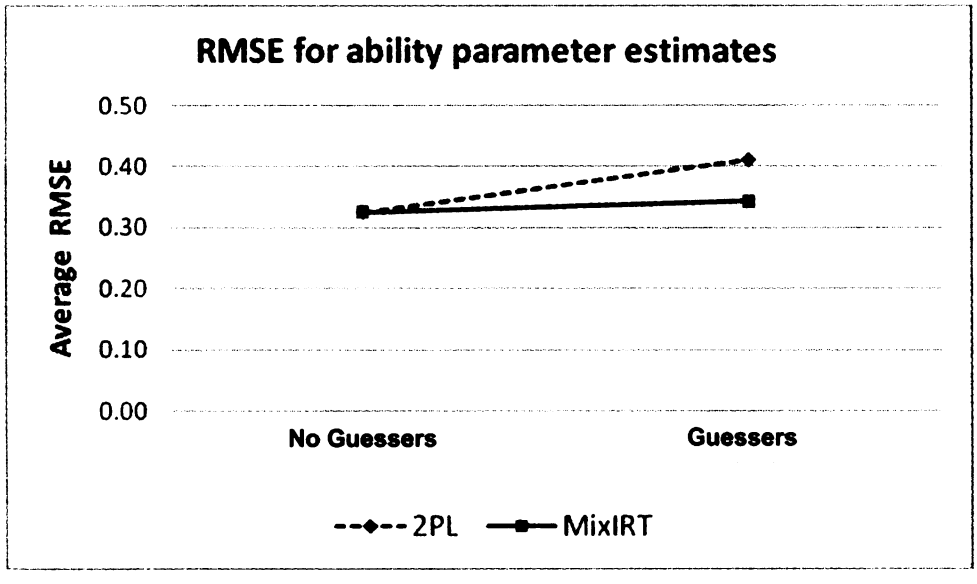


Figure 4.18 RMSE of ability parameter estimates in MixIRT-A model

Table 4.13 Correlation of Ability Parameter Estimates in MixIRT-A Model

IRT Model	No Guessers	Guessers
2PL	0.942	0.917
MixIRT	0.942	0.929

As mentioned earlier, classification accuracy is an important criterion for evaluating the degree to which the proposed MixIRT-A model accurately classifies examinees into their true (simulated) class or group. Table 4.14 provides weighted and unweighted classification accuracy results for the MixIRT-A model. The results indicate that this model correctly identified 63.12% of guessers. This indicates a lack of power in identifying the guessers. Also, misclassifications occurred from using this model. For the non-guesser class, 7.52% were incorrectly classified as guessers. Similarly, even for a sample with no guessers, the model incorrectly classified 3% of the examinees as guessers.

Table 4.14 Classification Accuracy of MixIRT-A Model

	True Class (Group*)	N	Average Estimated Guesser %	Classification Accuracy %	Weighted** Classification Accuracy %
Guessing Allowed	NG	748	7.52	92.48	
	G	252	63.12	63.12	85.08
No-Guessing	NG	1000	3.00	97.00	97.00
	G	0	NA	NA	

*NG=*Non-guessers*, G=*Guessers*

** Weighted by sample size

4.5 Results from Empirical Data Analysis

To address the fourth research question for which the goal was to investigate the impact of excluding aberrant item responses (from *guessers*) in proficiency level classification, real data from a statewide mathematics assessment was used. Since guessing behavior can only occur on multiple-choice items, the analyses were conducted on examinee responses to 54 multiple choice items. Because of extensive MCMC computational time, only two randomly selected samples of size 1000 were used in a cross-validation. The first sample is referred as a training sample and the second sample is referred as a validation sample. First, the results based on the random guessing model (MixIRT-R) are presented in section 4.5.1. However, the analysis was also carried out using the MixIRT model with ability-based guessing (MixIRT-A) so as to compare the classification of simulees into *guessers* and *non-guessers*. These results are presented in section 4.5.2.

4.5.1 Results Based on the Random Guessing Model

Tables 4.15 and 4.16 present sample WinBUGS output, particularly highlighting the estimates of class membership. The *node* in this table refers to the variable monitored in WinBUGS. In this output, PI[1] and PI[2] refer to classes or categories corresponding to *guesser* and *non-guesser* respectively. Interestingly, the estimates were similar for both samples, showing that about four to five percent of examinees were likely to belong to a *guesser* class in this particular assessment. The observation of 95% credible interval around the estimate and MC error being less than $1/20^{th}$ of the standard deviation indicates that these estimates are fairly precise.

Table 4.15 MixIRT-R Estimates for Training Sample

Node	Mean	Standard Deviation	MC error*	2.50%	Median	97.50%
PI[1]	0.044	0.008	< 0.001	0.029	0.044	0.062
PI[2]	0.956	0.008	< 0.000	0.938	0.956	0.971

*MC error: Monte carlo error

Table 4.16 MixIRT-R Estimates for Validation Sample

Node	Mean	Standard Deviation	MC error	2.50%	Median	97.50%
PI[1]	0.050	0.009	< 0.001	0.034	0.050	0.068
PI[2]	0.950	0.009	< 0.001	0.932	0.950	0.966

The estimates of guessing probability for each examinee also produced very similar results for both samples. Based on group membership estimate for each examinee, the numbers of *guesser* identified by the MixIRT-R model were 40 and 41 in training and validation samples respectively.

Three θ scale cut-scores were used for categorizing examinees into four proficiency levels based on values of -1.08, -0.53, and 0.39. As mentioned in the previous chapter, these cut-scores were chosen in such a way that the proportion of examinees into each proficiency levels in the current sample matched with that obtained from the actual statewide assessment. Therefore, in order to evaluate the impact of removing guessers from parameter estimation, the *guessers* identified by the MixIRT-R model were removed from the sample and the model parameters were estimated again. The results presented below are summarized for the same number of examinees, i.e. only *non-guessers* before and after removing *guessers* from the calibration.

Table 4.17 Distribution of Proficiency levels in Original and Modified Training Sample

	Original proficiency level		Modified proficiency level	
	Frequency	Percent	Frequency	Percent
Advanced	280	29.17	281	29.27
Proficient	373	38.85	367	38.23
Basic	243	25.31	234	24.38
Below Basic	64	6.67	78	8.13
Total	960	100	960	100

A closer look to these results does not indicate any noticeable differences in proficiency levels between the proportion of examinees before and after removing the guessers identified by the MixIRT-R model. For example, the percentage of examinees that were classified as proficient (*proficient* or *advanced*) changed slightly from 68.02 to 67.50.

Table 4.18 summarizes the distribution of proficiency levels for validation sample. The results from this sample were fairly similar to those obtained for training sample. There was a small difference between original and modified examinee classification as proficient (*proficient* or *advanced*) as indicated by changes in proficiency level from 68.20% to 68.30%.

Table 4.18 Distribution of Proficiency levels in Original and Modified Validation Sample

	Original proficiency level		Modified proficiency level	
	Frequency	Percent	Frequency	Percent
Advanced	281	29.30	283	29.51
Proficient	373	38.89	372	38.79
Basic	240	25.03	211	22.00
Below Basic	65	6.78	93	9.70
Total	959	100	959	100

Testing statistical significance of these differences would provide useful information for evaluating the meaningfulness of sample differences. As noted in Chapter 3, one way of comparing ability parameter frequency distributions was to use the Kolmogorov-Smirnov test. In addition to this test, a chi-square test was performed in order to test the statistical significance of the differences between proficiency levels classified by two samples. Table 4.19 provides Kolmogorov-Smirnov test results.

Table 4.19 Test Statistics from Two-sample Kolmogorov-Smirnov Test

		Training sample	Validation sample
		θ-Distributions	θ-Distributions
Most Extreme Differences	Kolmogorov-Smirnov Z	0.456	0.708
	Asymp. Sig. (2-tailed)	0.985	0.698

Kolmogorov-Smirnov test results from Table 4.19 suggest that the difference between two distributions is not statistically significant. Similarly, the results from the chi-square test suggest that the difference for original and modified proficiency levels is not significant for training sample (χ^2=1.60, df=3, p=0.66). The chi-square test also suggested no significant difference for validation sample (χ^2=6.83, df=3, p=0.08).

In an attempt to map the characteristics of the examinees classified into the *guesser* class from this analysis, no specific conclusions could be made in terms of gender and ethnicity. The only variable that seemed related with guessing was *economic disadvantage*(ED), a measure of socio-economic status, operationalized by *free or reduced lunch*. That is, ED=1 were more likely to be classified as guessers than ED=0.

4.5.2. Results Based on the Ability-based Guessing Model

The results based on the MixIRT-A are presented for both training and validation samples. This model identified that 7% of examinees were guessers for training sample, and 10% of the examinees were guessers for validation sample. Interestingly, among those 70 examinees for training sample and 100 examinees for validation sample that were classified as guessers by this model, 36 for training sample and 37 for validation sample were also classified as guessers by the previous model (MixIRT-R). This result is presented in Figure 4.19.

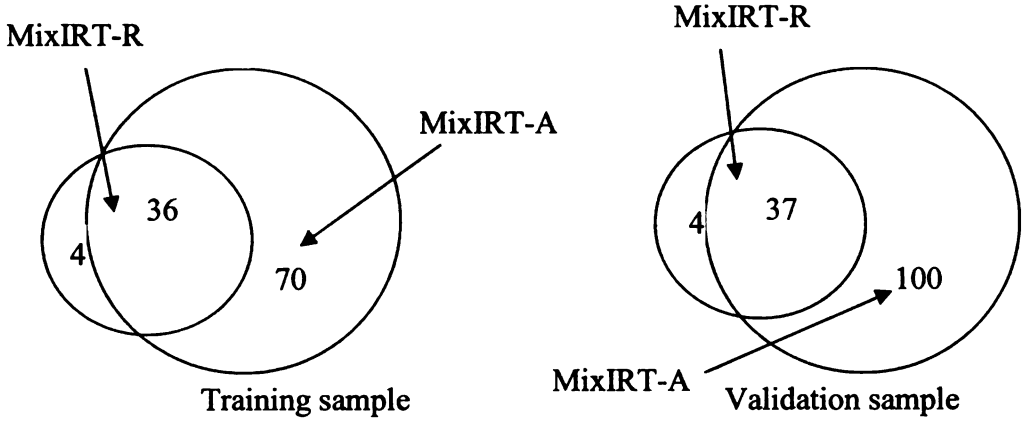


Figure 4.19 Number of examinees identified as guessers in training and validation sample

Table 4.20 presents the distribution of proficiency levels in the original and the modified training sample. Interestingly, the proficiency level for those who were proficient (*proficient* or *advanced*) has decreased from 68.17 (original sample) to 63.87 (modified sample). This shows a large change in the proficiency level.

Table 4.20 Distribution of Proficiency levels in Original and Modified Training Sample

	Original proficiency level		Modified proficiency level	
	Frequency	Percent	Frequency	Percent
Advanced	262	28.17	240	25.81
Proficient	372	40.00	354	38.06
Basic	233	25.05	240	25.81
Below Basic	63	6.77	96	10.32
Total	930	100	930	100

Table 4.21 presents the distribution of proficiency levels in original and modified sample for validation sample. Interestingly, the proficiency level for those who were proficient (*proficient* or *advanced*) has decreased from 68.11 (original sample) to 61.56 (modified sample). This also shows a large change in the proficiency level.

Table 4.21 Distribution of Proficiency levels in Original and Modified Validation Sample

	Original proficiency level		Modified proficiency level	
	Frequency	Percent	Frequency	Percent
Advanced	259	28.78	215	23.89
Proficient	354	39.33	339	37.67
Basic	226	25.11	253	28.11
Below Basic	61	6.78	93	10.33
Total	900	100	900	100

Table 4.22 Test statistics from Two-sample Kolmogorov-Smirnov Test

		Training sample	Validation sample
		θ-Distributions	θ-Distributions
Most Extreme Differences	Kolmogorov-Smirnov Z	1.322	1.721
	Asymp. Sig. (2-tailed)	0.061	0.005

The statistical test of differences between two distributions in original and modified sample suggested mixed findings for statistical significance at the α-level of 0.05. For the training sample, the Kolmogorov-Smirnov Z shows non-significant (Z=1.322, p=0.061). However, for the validation sample, it shows a significant difference as indicated Z value of 1.72 ($p < 0.05$).

In the chi-square test, the differences in proficiency levels (cell frequencies) between original and modified sample were statistically significant for both training and validation samples. For example, the chi-square test statistics of χ^2=8.363, df=3, p=0.039 indicate the statistically significant difference between cell frequencies for training sample and χ^2=12.58, df=3, p=0.006 indicate the statistically significant difference between cell frequencies for validation sample.

The next chapter provides discussion and conclusions for this study. It summarizes the results, interprets those findings, and lists some implications of those results.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

The primary goal of this study was to explore the effectiveness of mixture IRT (MixIRT) models in estimating the differential performance of latent classes in a sample (i.e., sample heterogeneity). The variables (e.g, *guessers* or *non-guessers*) used for classifying examinees are referred to as sources of heterogeneity. When test-taking heterogeneity sources are unobservable (e.g., examinees' tendency to guess), and if their group membership has to be inferred from the data, *unobserved test-taking heterogeneity* is said to exist.

Therefore, in this study, the MixIRT model was used to investigate different examinee test-taking behaviors through a simulation study that varied (a) sample size, (b) test length, and (c) proportion of guessing. These factors were selected because these were thought to be useful in many testing applications like item pool design or IRT-based test bank development and pre-equating where precision of parameter estimation is paramount. Furthermore, varying these factors allowed the extent to which differing degrees of test-taking heterogeneity influence model parameter estimation to be studied, particularly for different test lengths and sample sizes.

Given that MixIRT models are an extension of IRT models, their parameter estimation is complicated by the intractability of mathematical forms when trying to use frequentist techniques. Therefore, Bayesian estimation was used instead because it can handle high-dimensional problems and the distributions of parameters, can be explained

regardless of the forms of the distributions of the likelihood and the parameters. Through a simulation study, the precision of parameter estimation was evaluated in the MixIRT model for various realistic testing factors. As mentioned in Chapter Three, this study used two forms of MixIRT model to incorporate different guessing strategies, viz. MixIRT-R and MixIRT-A.

Considering the extensive computational time required for the MCMC procedures of Bayesian methods that were used, only two levels of test length and sample size were considered. Since the impact of unobserved test-taking heterogeneity, represented in this study as a proportion of guessers, on model parameter estimation was the primary factor of interest, the proportion of guessers per sample was varied. Two percentages of guessing were used to represent 5% and 10% of the total examinees as guessers. The data with no guessers, represented as *0% guessing proportion,* was used as a baseline to compare the results.

Another purpose of this study was to compare the parameter estimation accuracy of the MixIRT model to two commonly used IRT models: the 2PL and 3PL models. A parameter recovery study was used for conducting the aforementioned comparison. This comparison was carried out by varying the three estimation models (i.e., 2PL, 3PL, and MixIRT) for all the study factors in a fully crossed design. The precision of parameter recovery was evaluated based on three commonly used evaluation criteria: bias, RMSE, and Pearson correlation. The interpretations of the results were based on both numeric and graphic representations.

The study's third objective was to evaluate the accuracy of MixIRT Bayesian estimation in identifying guessers when there were *guessers* in a sample. For this

purpose, the MixIRT model estimated the probability that each examinee belonged to latent classes of either *guessers* or *non-guessers*. And the model's classification accuracy, which indicates the extent to which simulees are correctly categorized as *guessers* or *non-guessers*, was evaluated. As noted earlier, the probability of an examinee likely to be a *guesser* was estimated from the item response pattern of the examinee and the probability was actually based on the average over a large number of MCMC iteration. In this study, the examinee was classified as a *guesser* if that probability was equal to or greater than 0.5.

The study's final purpose was to investigate the impact of excluding aberrant guessing responses in examinee proficiency level classification. In other words, the ability continuum was divided into four different levels so that the impact could be studied in terms of proficiency classification. For proficiency classification, real data was used as a further illustration of the MixIRT model's usefulness. This goal has potential for contributing to the better understanding of issues pertaining to cut-scores variation and its policy implications.

It is important to clarify that this study does not suggest guessing is a bad thing from a student's perspective, especially in circumstances such as when there is no penalty for guessing and when examinees run out of time. However, from the measurement or psychometric point of view, guessing introduces construct-irrelevant variance, which is a major concern in validity studies. Therefore, the objective of this study was to document the impact of guessing on parameter estimation thereby influencing proficiency level classification. In simple terms, the practical example illustrated in this dissertation was similar to using a correction for guessing to get the corrected distribution. Therefore, the

goal was to illustrate how the proposed mixture modeling approach has potential to address this very important issue encountered in many large scale assessments.

5.1 Interpretations of the Results

5.1.1 Results from Parameter Recovery Study

As noted previously, one of this study's major goals was to evaluate the accuracy of parameter estimates by comparing them to true (simulated) parameters. The results presented both numerically in Tables 4.3 and 4.4 and graphically in Figures 4.2 to 4.5 show that both bias and RMSE values for discrimination and difficulty parameters are generally lower in MixIRT-R model estimation as compared to those obtained from the 2PL model.

When no guessers were present in the sample, the bias and RMSE values were similar in both MixIRT-R and 2PL models. The lower values of these indices show that parameters are estimated reasonably well when no aberrant responses are present in the data. However, bias and RMSE values tended to be higher for both models when the proportion of guessers in the sample increased to 5% and 10%. This suggests that even in presence of 10% guessers in the sample, the aberrant responses have a huge impact on precision of item parameter estimation. Since commonly used IRT models (e.g., 1PL, 2PL, 3PL) are not designed to handle the test-taking heterogeneity, alternate modeling approaches are necessary. A mixture model provides such avenues by allowing different latent classes to have their own set of model parameters.

One of the primary objectives of varying study factors like test lengh and sample size was to evaluate their capacity to recover stipulated item and person parameters. No clear interpretation could be drawn from the available evidence about the impact of test

length on bias and RMSE. Moreover, as other studies have also shown, the larger sample size resulted in smaller bias and RMSE. This was, however, not the case for the 2PL model when test length was 50 and the sample size was 2000. These findings play an important role in judging the quality of IRT-based test banks and pre-equating used in large scale assessments.

The average correlations between true (simulated) and estimated item parameters, presented in Table 4.5 and Figures 4.6 to 4.9 show the recovery of item discrimination and item difficulty parameters. Stronger correlations were associated with larger sample sizes for both 2PL and MixIRT-R models. This finding was also consistent with the literature on IRT parameter recovery.

The impact of guessing was profound in recovery of item discrimination parameters for the 2PL model. For example, the correlation between true and estimated *a* parameters decreased from 0.877 to 0.807 with the use of 2PL model when the proportion of guessers increased from 5% to 10%. The correlations were similar for both 2PL and MixIRT-R models when no guessers were included in the sample. This suggests that when unobserved test-taking heterogeneity is absent (i.e., no guessers are present in the sample), it may not be necessary to use the complex models like the MixIRT. Nevertheless, this situation may not be practical in most situations as guessing is widely known to occur in many large scale assessments.

Overall, difficulty parameters had better recovery than discrimination parameters. This is consistent with the findings from earlier research, which showed that discrimination parameters are usually more poorly estimated than the difficulty parameters.

Person parameter recovery results are summarized in Tables 4.6 and 4.7. Sample plots of the parameter recovery results are also presented in Figures 4.14 and 4.15. The 2PL and MixIRT-R bias and RMSE values were similar, suggesting that ability parameter recovery was fairly similar for both models. However, among the three models compared in this study, the 3PL model performed the worst as indicated by large bias and large RMSE. One possible reason for this poor performance could be a result of the types of guessing behavior introduced in this simulation. That is, guessing is defined as examinee behavior and estimated as a person parameter using a probabilistic model.

Generally, in the IRT framework, studies that use the 3PL model simulate data by associating guessing as a parameter associated with the items indicated by the *c* parameter. In addition, *c*-parameters are often recovered very poorly (Martin et al., 2006; Pelton, 2002), because these are generally estimated as lower asymptotes based on fewer number of examinees. In this study, the model fit index *Deviance Information Criteria* (Spiegelhalter, Best, Carlin, & van der Linde, 2002) showed that the fit of 2PL model was better than that of 3PL model even when 10% guessing proportion was present. Finally, the true (simulated) parameters were generated based on the 2PL model and introduction of *guessers* might have noticeable impact that could not be captured by the 3PL model.

Furthermore, based on the correlations between estimated and simulated ability parameters, the results were fairly similar in all three models and the magnitude of correlation was generally strong. This indicates that the influence of unobserved test-taking heterogeneity was more noticeable in item parameter estimation than the person parameter estimation. This finding suggests that the proposed mixture modeling approach

is more appropriate in applications where precise estimation of item parameter is paramount, such as pre-equating and IRT-based item banking or item pool.

5.1.2 Results on Classification Accuracy

To investigate the accuracy of Bayesian MixIRT model estimation in correctly classifying *guessers*, this study evaluated the results based on an index called the classification accuracy. Table 4.8 shows the classification accuracy when using the MixIRT model. The classification accuracy was over 98% for the membership to the *non-guesser* group and approximately 90% for the membership to the *guesser* group. In terms of weighted classification accuracy, the MixIRT model performed better in classifying examinees into the group where they belonged. This was reflected by the accuracy of 96.92% or higher in all simulated conditions. The classification accuracy was 100% for the conditions when there were no *guessers* in the sample. This finding suggests that the MixIRT models can be used even in absence of unobserved test-taking heterogeneity. However, due to the complexity of the mixture models and the costs associated estimating a large number of parameters, there is no advantage of using the MixIRT model when no guessers are present.

5.1.3 Results from Empirical Study

Finally, results from the real data example are presented in Tables 4.11 and 4.12. The MixIRT-R model identified that nearly 5% of examinees were likely to be guessers in this sample. The precision of these estimates are reflected in a 95% credible interval

around the estimate and the fact that MC error is less than $1/20^{th}$ of the standard deviation.

The impact of excluding *guessers* in parameter estimation was also expressed in terms of classification into proficiency levels. As mentioned in Chapter 3, this study used four proficiency levels: *Advanced*, *Proficient*, *Basic*, and *Below Basic*, which are commonly used in current test-based accountability system under NCLB. To evaluate the degree to which the proficiency level classifications differ between the two samples, with or without the guessers, a chi-square test was performed to compare whether the proportions of student in proficiency levels are different between two samples. In addition, the distributions of two ability (θ) estimates using a two-sample Kolmogorov-Smirnov test show that the differences were not statistically significant for both samples. This suggests that the specified MixIRT model did not find guessing as a potential cause of observed difference in proficiency classification for this assessment. Furthermore, the influence of a small proportion of guessers in the sample, i.e., less than 5%, did not have much influence on parameter estimation and decisions regarding proficiency level classification. It might be possible that, for an assessment where more examinees are engaged in guessing, the impact on parameter estimation, as well as proficiency level classification, could be noticeable.

The impact of guessing was noticeable in analysis using the MixIRT-A model. This model identified 7% of examinees as *guessers* in the training sample, and 10% in the validation sample. Interestingly, among those 70 examinees for training sample and 100 examinees for validation sample that were classified as guessers by this model, 36 and 37 for training and validation samples were also classified as guessers by the previous

model, i.e. MixIRT-R. This shows that the two models are related. Naturally, the ability-based guessing model is expected to identify more guessers than the random-guessing model.

As earlier the impact of excluding guessers from the sample was evaluated by finding the differences in proficiency level and ability distribution of examinees before and after removing the guessers from the calibration. Interestingly, the proficiency level for those who were proficient (*proficient* or *advanced*) was changed from the original to the modified sample for both training and validation samples. For example, the percent of proficient was changed from 68.17 to 63.87 in training sample showing a noticeable impact of removing guessers from the calibration. Similarly for validation sample, Table 4.21 showed that the proficiency level for those who were proficient (*proficient* or *advanced*) had changed from original to modified sample. The changes in proficiency level from 68.11% to 61.56% also indicate a noticeable impact on proficiency level classification.

In terms of inferential statistics, the statistical test of differences between the two distributions in the original and modified samples had mixed findings at a statistical significance at alpha level of 0.05. In other words, the Kolmogorov-Smirnov Z shows non-significant result for the training sample but shows a significant difference for the validation sample. For the chi-square test, the differences in proficiency levels between the original and modified samples were statistically significant for both samples. This may have a large ramification from a policy perspective because even a few percent changes in proficiency level receive attention by teachers, school administrators, and policy makers. In this context, it may be prudent to decide to locate the appropriate cut-

score on the continuum where few examinees are situated so that a change in the cut-score would result in unnoticeable changes in proficiency classification. The change in students' proficiency estimates are also of interest to a wider audience that includes parents, teachers, school administrators, educational researchers, and policy makers. This puts unique responsibilities on educational researchers and psychometricians to answer the question of whether the changes in student's proficiency estimates are associated with actual improvement in their ability or due to a measurement or scaling issues.

5.2 Study Limitations

Like any simulation study, there are also questions regarding the generalizability of findings to real testing situations. Utmost care was taken to ensure that the simulated conditions match with practical settings. However, due to limited flexibility of modeling in the program used for MCMC sampling in this study, i.e. WinBUGS, it was not possible to take full advantage of Bayesian inference. For example, a user has limited control over sampling procedures implemented in WinBUGS. The DIC value was not possible to compute for the mixture model in order to compare the model fit statistics. Therefore, the model fit evaluation of the mixture model was limited to a likelihood ratio test.

The findings documented in this study were based on 15 replications. This may also raise a question about the generalizability of the findings. However, this decision was made due to the slow performance of WinBUGS. It should be noted that parameter estimation using Gibbs sampling requires a substantial amount of time, especially when it is estimated using WinBUGS. In this study, the required computing time for each dataset

varied anywhere from 1 hour to 6 hours of computer time (with Intel Centrino Duo processor 1.66GHZ, 2 GB RAM) depending upon the sample size and test length. Therefore, this study used a limited number of levels within each of the simulated factors.

The MCMC estimation were performed using WinBUGS. Therefore, it should be noted that the results obtained may not just be due to the theoretical differences between models and study factors, but may also be related to how the software implements the MCMC methods. In other words, use of alternative software or methods of estimation could potentially lead to different results. Thus, comparing the results obtained from the Bayesian estimation method implemented in WinBUGS with other programs such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) or Mplus (Muthén & Muthén, 1998-2007) may provide an additional perspective on this issue.

In a real-data application, this study classified the examinees into latent classes of *guessers* and *non-guessers* based on the probability of each examinees being classified into each class. The recommendation of this study to delete the guessers from the calibration to improve the parameter estimation may not always be realistic in many large scale assessments (e.g., state large scale assessment) because states are required to report the scores for each examinee.

5.3 Implications

There are several possible implications of this dissertation. First, this dissertation explored the effectiveness of using MixIRT models to estimate the differential performance of latent classes in a sample i.e., sample heterogeneity. By studying the impact of various study factors like sample size, test length, and proportion of guessing

on parameter recovery, it provides useful information for various testing applications like item pool design or IRT-based test bank development and pre-equating. This study also provides some direction on identifying aberrant item responses in any large scale assessment and mapping the profile of guessers.

This study focused on an important issue of psychometrics, i.e. parameter estimation. If parameters are not well estimated, the proportion in Adequate Yearly Progress (AYP) category will not be accurately reported. This study has the potential to increase our understanding of challenges and conditions in the modeling of complex behavioral phenomenon, such as test-taking behaviors. Furthermore, illustrative use of WinBUGS may encourage researchers and practitioners to utilize Bayesian methods for investigations of alternate modeling strategies when their data do not fit a single IRT model. Finally, this dissertation also provided a substantive and policy-relevant illustration of ignoring test-taking heterogeneity, especially by showing how simplistic applications of 2PL and 3PL models could adversely impact not only the examinees but also schools, teachers, policymakers.

5.4 Future Directions

There are several possible future directions for this study. First, a future study could model complex guessing strategies by taking interaction of ability, item difficulty, and item location into account. Although the 3PL model performed less well than the 2PL in this study potentially due to the guessing simulation favoring the 2PL, in practical situations, we may not know which model would perform better. Therefore, a comparative study of the 2PL and 3PL using real data may provide some useful findings.

Similarly, the estimations from this study could be compared with those from other estimation programs, such as BILOG-MG (Zimowski et al., 2003), Mplus (Muthén & Muthén, 1998-2007), and mdltm (von Davier, 2005). Although BILOG-MG is not intended to model unobserved test-taking heterogeneity, the recovery of parameter could still be compared especially for the sample with no-guesser (0% proportion of guesser in this study) and the sample after removing guessers. Also, comparing the MixIRT model parameter estimation from WinBUGS with that from M-Plus or mdltm might provide an additional perspective on the direction that could be undertaken if the MixIRT model approach has to be realized in the practical situations.

In terms of handling the number of classes in the mixture model, the present study was limited to two latent classes. Future studies could also explore such investigation using more than two classes. The complexity of mixture modeling is further increased by simulating the mixtures of one-parameter and 2-parameter IRT models, or even mixtures of unidimensional and multidimensional IRT models. Such mixture IRT modeling has potential to provide useful information for applications such as sub-score reporting and cognitive diagnostic modeling. Future studies could investigate the practicality of estimating such complex models in the mixture IRT framework.

As indicated in several earlier studies using WinBUGS, use of low level programming languages such as FORTRAN or C++ to implement Gibbs sampling may provide more flexibility in addition to reduced computational time. Some of the limited flexibility of modeling in this study should not be attributed to the Bayesian estimation so much as to the estimation tool used in this study, i.e. WinBUGS. Therefore, we may gain

some modeling and computational efficiency while moving in the direction of using a low level programming language.

5.5 Summary of the Findings and Conclusions

In summary, this study shows that the MixIRT model can precisely recover the model parameter. It also found that ignored unobserved test-taking heterogeneity, like the presence of guessers in a sample in this study, had a noticeable impact on the precision of recovery of both item and ability parameters. The item parameters were estimated more precisely in MixIRT as compared to 2PL model. Finally, the mixture IRT model classified examinees into *guessers* and *non-guessers* reasonably well. The impact of guessing on ability estimation was not severe when the percentage of guessers was low, i.e. less than 5%. However, when the proportion of guesser was higher, say 7% to 10%, the impact was noticeable as indicated by significant changes in examinees classified into proficiency levels.

This study investigated an important psychometric issue in large scale assessment, such as modeling unobserved test-taking heterogeneity, using IRT mixture model. It identified the guessers by estimating the probability based on response pattern of examinees. This study also documented the impact of excluding the guessers from the calibration to improve the parameter estimation, which has a large impact on improving the quality of IRT-based item banking and the inferences drawn from the tests assembled using the item pool. Since states are required to report the scores for each examinee, the recommendation suggested by the results of this study to delete the guessers from the calibration to improve the parameter estimation may not always be practical in many

large scale assessments. However, this study suggests that the proposed mixture modeling approach can be applied in many large scale assessment such as IRT-based item banking to improve the quality of pre-equating and any inferences drawn from the item parameter estimation.

This dissertation explored a psychometric perspective of modeling guessing as a person characteristic rather than associating it with the item property as is commonly done with the three-parameter logistic model. Finally, use of real data and illustration of the MixIRT model's usefulness in documenting the changes in proficiency level classifications has potential for improving the understanding of issues pertaining to cut-scores variation and its policy implications.

APPENDICES

List of Appendix

A. WinBUGS CODE FOR MixIRT model
B. FIGURES FOR EVALUATING CONVERGENCE OF THE ESTIMATES
C. ADDITIONAL TABLES
D. ADDITIONAL PLOTS

APPENDIX A

```
#####################################
# Mixture IRT model
#####################################

model
{
for (i in 1:N) {
                for (j in 1:J) {
                        p[i,j] <- (2-G[i])/5 + (G[i]-1) * 1/ (1+exp(-(a[j]*(theta[i]-b[j])))) ;

                                r[i,j] ~ dbern(p[i,j]) ;
                                }
                        G[i] ~dcat(PI[]);
                                pg[i] <- equals(G[i],1);
                                        #probability of being in a guesser class
                        }

        # priors

        PI[1:2] ~ ddirch(alpha[]);

                for (j in 1:I) {
                a[j] ~ dnorm(1,2) I(0,); #Truncated Normal
                # a[j] ~ dlnorm(0,2);  #Log Normal
                b[j]~ dnorm(0, 1) ;
                                }

        for (i in 1:N) {
                theta[i] ~dnorm(0, taut) ;   #prior for ability parameter
                }
                taut ~dgamma(0.01,0.01);

        }
```

Appendix A.2 WinBUGS code for Model 2 (MixIRT model with ability based guessing)

```
model
{
for (i in 1:N)    # N is the number of examinees
        {
        for (j in 1:J)   # J is the number of items
        {

logit(p[i,j]) <- (a[j]*(theta[i]-b[j]))-(alpha[i]-1)*step(b[j]-theta[i]-delta[i])*(a[j]*(theta[i]-
b[j])+1.4);

check[i,j] <- (alpha[i]-1) * step(b[j]-theta[i]-delta[i]);

# 1.4 value is given instead of estimating c, because logit (-1.4) = 0.20
# step function (b[j]-theta[i]-delta[i]) =1 if an item j is difficult for an examinee i with
threshold delta
# alpha estimates the group membership

                        r[i,j] ~ dbern(p[i,j]) ;
                                }
                                sumcheck[i] <- sum(check[i,1:J]);

                                check2[i] <- sumcheck[i];

                                alpha[i] ~dcat(PI[]); # group membership is categorical
                                                }

        #priors
        for (j in 1:J) {
                b[j]~ dnorm(0, 1) ; # Normal prior for difficulty parameter
                a[j] ~ dnorm(1,2) I(0,);  # Truncated normal for discrimination parameter

        }

                for (i in 1:N) {
                theta[i] ~dnorm(0, taut) ;  #prior for ability parameter
                delta[i] ~ dnorm(0,10); #threshold for different degree of guessing
                }
                taut ~dgamma(0.01,0.01);  # hyper-parameter for precision

        PI[1] ~ dbeta(1,1);
        PI[2] <- 1.0- PI[1];
}
```

APPENDIX B

FIGURES FOR EVALUATING CONVERGENCE OF THE MCMC METHODS

Figure B.1 Convergence Diagnostic Plots for a difficulty parameter of a randomly selected item [*True b =2.08, Estimated b =2.125]*

Figure B.1a BGR plot for difficulty parameter estimate

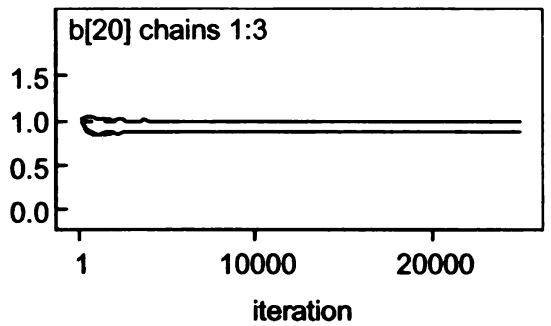


Figure B.1b. History plot for difficulty parameter estimate

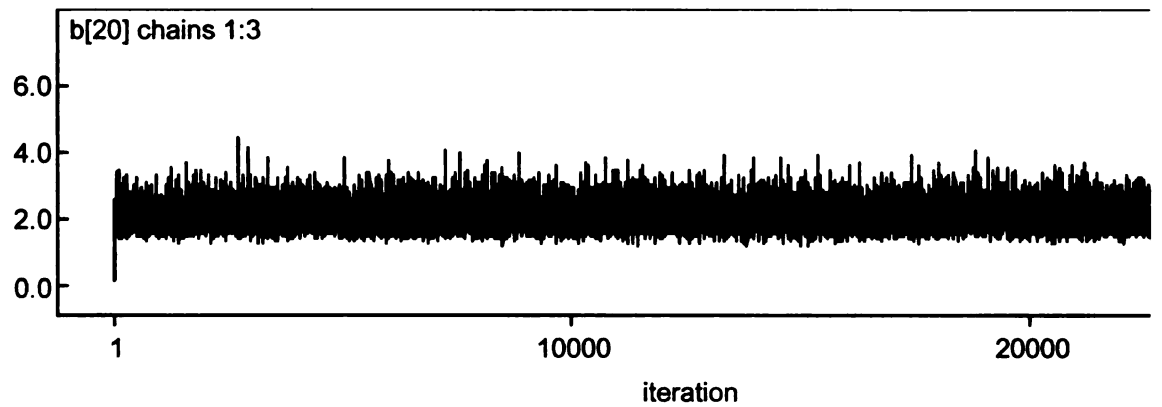


Figure B.1c. Autocorrelation plot for difficulty parameter estimate

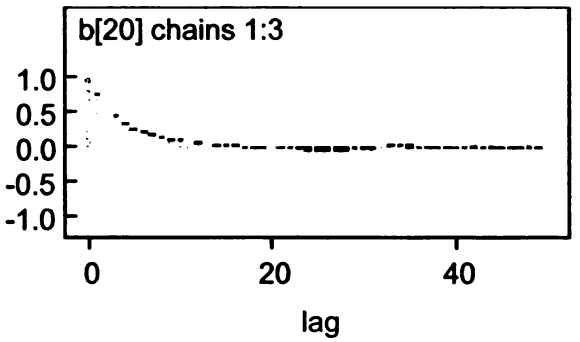


Figure B.1d. Density plot for difficulty parameter estimate

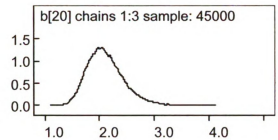


Figure B.2 Convergence Diagnostic Plots for an ability parameter of a randomly selected person *[True theta=1.365, Estimated=0.862]*

Figure B.2a BGR plot for ability parameter estimate

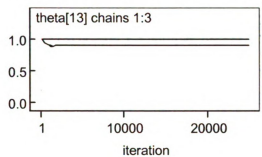


Figure B.2b. History plot for ability parameter estimate

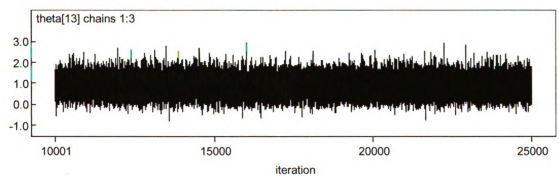


Figure B.2c. Autocorrelation plot for ability parameter estimate

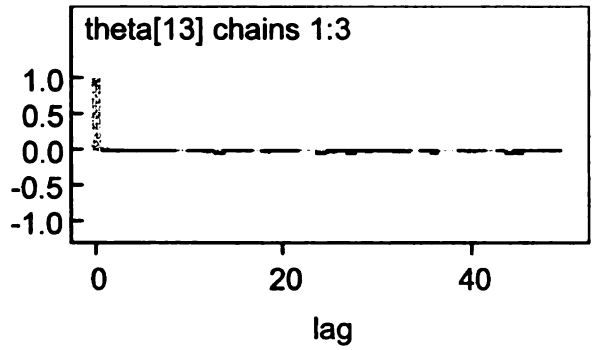
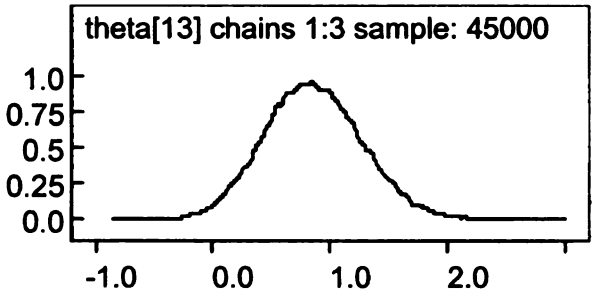


Figure B.2d. Density plot for ability parameter estimate



Appendix C. ADDITIONAL TABLES

Table C.1 Simulated Item Parameters (Test Length=25)

Item Number	Discrimination (a)	Difficulty (b)
1	0.98	0.17
2	1.22	0.15
3	0.97	0.23
4	1.17	-1.22
5	1.48	-1.90
6	0.99	-1.63
7	1.08	0.98
8	1.64	-0.29
9	0.85	-0.51
10	1.08	-0.18
11	1.49	0.14
12	0.78	0.72
13	0.75	-0.98
14	0.69	0.10
15	0.79	0.65
16	0.99	-0.85
17	0.59	-1.61
18	0.63	0.19
19	0.68	0.12
20	0.94	2.09
21	1.35	-0.04
22	0.99	-0.01
23	0.87	-0.41
24	0.98	0.97
25	1.76	0.09

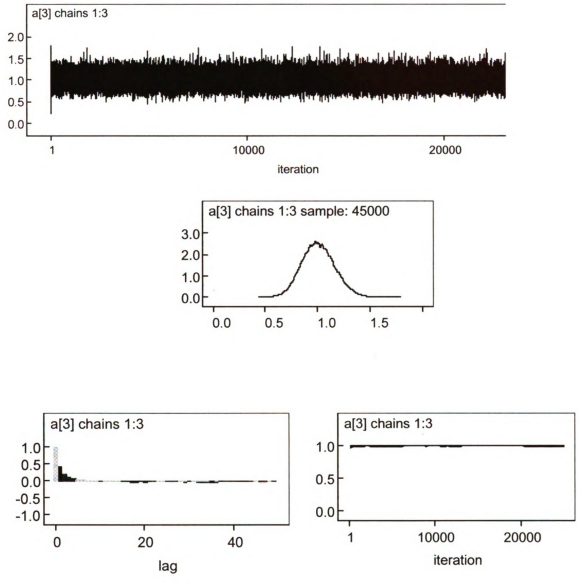
Table C.2 Simulated Item Parameters (Test Length=50)

Item Number	Discrimination (a)	Difficulty (b)	Item Number	Discrimination (a)	Difficulty (b)
1	0.70	0.53	26	1.19	0.19
2	1.17	-1.11	27	1.15	-0.28
3	0.72	-1.04	28	1.28	0.71
4	0.85	-0.48	29	1.37	1.26
5	0.96	-1.19	30	1.01	-0.89
6	1.06	0.45	31	0.92	-0.02
7	0.98	-0.22	32	1.06	-0.89
8	1.54	-1.03	33	1.28	0.60
9	1.50	-0.71	34	1.36	0.40
10	1.17	-0.79	35	0.87	-0.41
11	1.02	-0.06	36	1.26	-0.45
12	0.82	-1.93	37	0.96	-0.60
13	0.89	-0.56	38	0.92	0.74
14	0.70	-1.11	39	1.14	-0.20
15	0.91	0.26	40	1.26	-0.12
16	0.89	0.48	41	0.87	-1.01
17	1.55	0.12	42	0.69	0.25
18	1.21	-0.08	43	0.99	2.15
19	1.02	-2.11	44	1.15	0.23
20	0.93	0.02	45	1.02	-1.14
21	1.02	-0.70	46	1.48	0.20
22	1.21	-0.21	47	0.92	-1.15
23	1.75	-0.26	48	1.22	0.12
24	1.09	-0.93	49	0.82	-1.35
25	1.24	0.46	50	0.73	0.58

Table C.3 Simulated Item Parameters (Test Length=40)

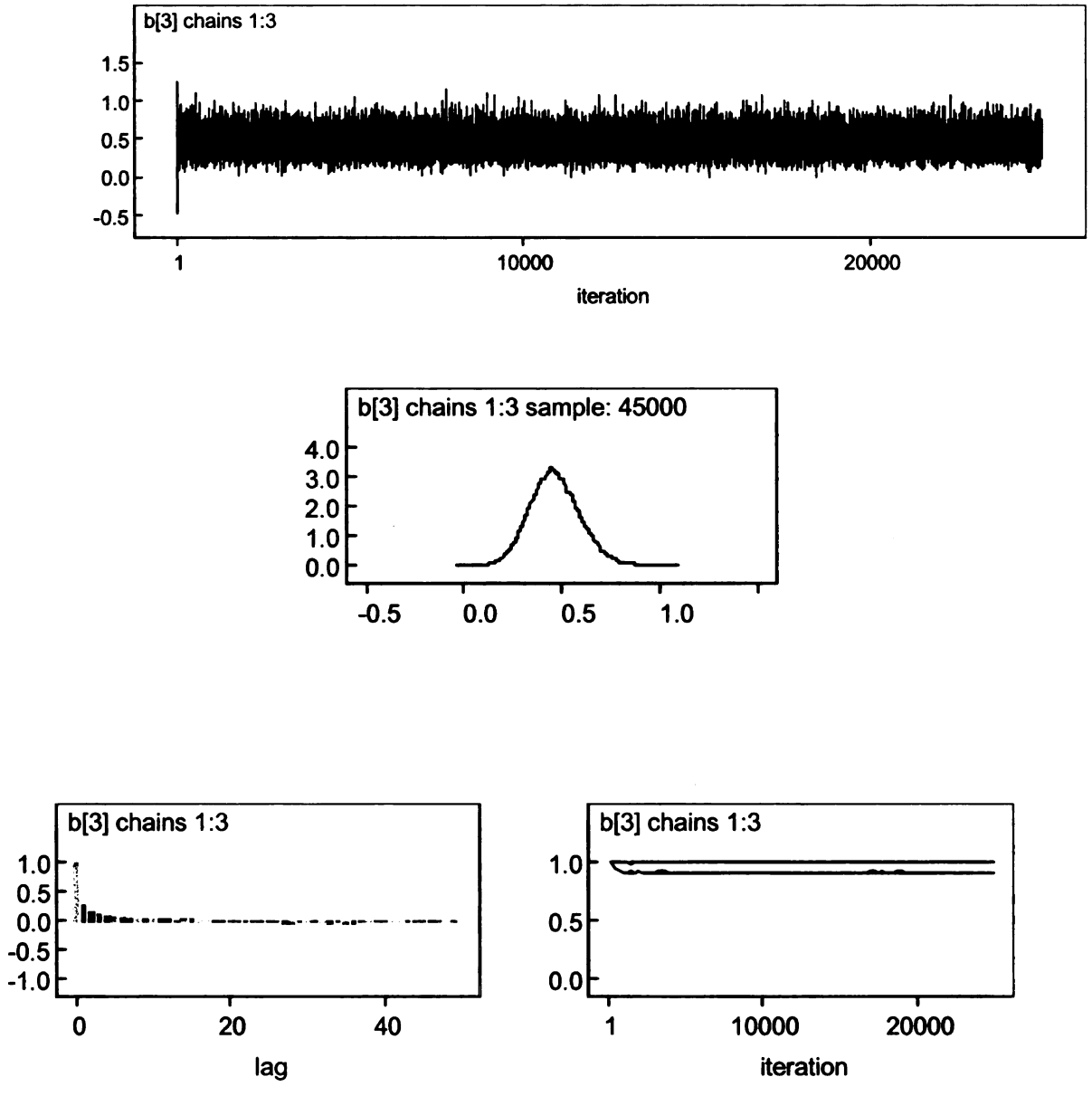
Item Number	Discrimination (a)	Difficulty (b)	Item Number	Discrimination (a)	Difficulty (b)
1	0.853	0.575	21	1.298	0.694
2	1.176	-0.226	22	0.785	1.945
3	1.227	-1.140	23	0.798	0.088
4	1.175	0.369	24	0.800	0.021
5	0.858	0.873	25	0.911	0.776
6	0.673	-0.835	26	0.633	-0.004
7	0.833	-2.274	27	1.281	1.128
8	0.844	0.797	28	0.832	1.406
9	1.026	0.061	29	1.334	0.708
10	1.231	-1.493	30	1.807	0.913
11	1.897	-0.491	31	1.093	0.323
12	0.999	0.935	32	0.889	0.153
13	0.974	-0.460	33	1.189	-0.555
14	0.926	-0.212	34	0.710	0.009
15	0.769	0.004	35	1.019	-0.884
16	1.135	-0.032	36	1.004	1.526
17	0.961	-0.404	37	0.951	-0.132
18	1.176	-0.926	38	0.814	-0.586
19	1.300	0.568	39	0.743	-0.793
20	0.687	0.583	40	0.985	0.715

Figure D.1 Convergence Diagnostic Plots for a discrimination parameter of a randomly selected item [*True a =0.966, Estimated a =1.008]*



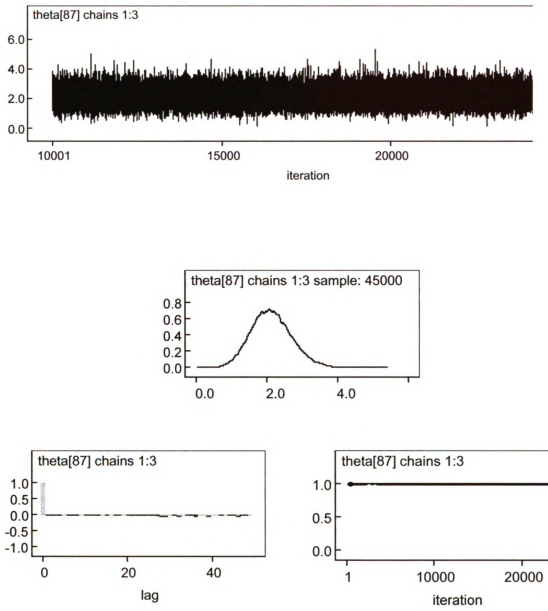
[From Top: History plot, Density plot, Autocorrelation plot (Left), BGR plot (Right)]

Figure D.2 Convergence Diagnostic Plots for a difficulty parameter estimate of a randomly selected item [*True b =0.2278, Estimated b =0.4627]*



[From Top: History plot, Density plot, Autocorrelation plot (Left), BGR plot (Right)]

Figure D.3 Convergence Diagnostic Plots for an ability parameter estimate of a randomly selected examinee *[True θ=2.545, Estimated θ=2.137]*



[From Top: History plot, Density plot, Autocorrelation plot (Left), BGR plot (Right)]

REFERENCES

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

Albert, J. H., & Ghosh, M. (2000). Item response modeling. In D. Dey, S. K. Ghosh & B. K. Mallick (Eds.), *Generalized Linear Models: A Bayesian Perspective* (pp. 173-193). New York: Addison-Wesley.

Ansari, A., Jedidi, K., & Dube, L. (2002). Heterogeneous factor analysis models: A Bayesian approach. *Psychometrika.*

Asparouhov, T., & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models*. Charlote, NC: Information Age Publishing.

Baker, F. B., & Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques*: CRC Press.

Bazan, J. L., Branco, M. D., & Bolfarine, H. (2006). A skew item response model. *Bayesian Analysis, 1*(4), 861-892.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models *Psychometrika, 66*(4), 541-561.

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*: Information Age Publishing.

Bock, R. D., & Zimowski, M. F. (1997). Multiple Group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 433-448). New York: Springer Verlag.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395-414.

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

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement, 30*(4), 277-291.

Cao, J., & Stokes, S. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika, 73*(2), 209-230.

Casella, G., & George, E. (1992). Explaining the Gibbs sampler. *The American Statistician, 46*(3), 167-174.

Cizek, G. J. (2001). *An overview of issues concerning cheating on large-scale tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, April 2001, Seatle, Washington.

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

Congdon, P. (2005). Markov Chain Monte Carlo and Bayesian statistics. In B. Everitt & D. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 3, pp. 1134-1143): Wiley.

Cowles, M. (2004). Review of WinBUGS 1.4. *American Statistician, 58*(4).

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

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of Royal Statistical Society, Series B, 39*, 1-38.

Draney, K., Wilson, M., Gluck, J., & Spiel, C. (2008). Mixture models in a developmental context In G. R. Hancock & K. Samuelsen (Eds.), *Advances in Latent Variable Mixture Models*. Charlote, NC: Information Age Publishing.

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

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). Retrieved October 21, 2008, from http://www.ncme.org/pubs/items/ITEMS_Mod_4.pdf

Frühwirth-Schnatter, S. (2006). *Finite mixture and markov switching models*: Springer.

Gelfand, A., Hills, S., Racine-Poon, A., & Smith, A. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association, 85*, 972-985.

Gelfand, A., & Smith, A. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association, 85*, 398-409.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-511.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721-741.

Gilks, W., Richardson, S., & Spiegelhalter, D. (Eds.). (1996). *Markov Chain Monte Carlo in practice*: Chapman & Hall/CRC.

Gill, J. (2002). *Bayesian methods (A Social and Behavioral Sciences Approach)*: Chapman & Hall/CRC.

Goldman, S. H., & Raju, N. S. (1986). Recovery of one- and two-parameter logistic item parameters: An empirical study. *Educational And Psychological Measurement, vol, 46*(1), 11-21.

Gonzales, P., Calsyn, C., Jocelyn, L., Mak, K., Kastberg, D., Arfeh, S., et al. (2000). *Pursuing excellence: Comparisons of international eighth-grade Mathematics and Science achievement from a U.S. perspective, 1995 and 1999* (No. NCES 2001-028): National Center for Educational Statistics.

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

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York: Macmillan.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer.

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (Eds.). (2002). *Making sense of test-based accountability in education*. Santa Monica, CA: RAND.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*, 97-109.

Hulin, C., Lissak, R., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characterstic curves: A monte carlo study. *Applied Psychological Measurement, 6*(3), 249-260.

Johnson, E. G. (1992). The design of the national assessment of educational progress. *Journal of Educational Measurement, 22*, 95-110.

Johnson, V. E. (1997). On Bayesian analysis of multirater ordinal data: An application to automated essay grading. *Journal of the American Statistical Association, 91*, 42-51.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal Data Modeling*. New York: Springer-Verlag.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *The Annals of Mathematical Statistics, 27*(4), 887-906.

Kim, J. S., & Bolt, D. M. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice, 26*(4), 38-51.

Kim, S. H., & Cohen, A. S. (1998). *An evaluation of a Markov Chain Monte Carlo method for the two-parameter logistic models.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kolmogorov, A. N. (1933). On the empirical determination of a distribution function. *4*, 83-91.

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

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

Lemke, M., & Gonzales, P. (2006). *U.S. student and adult performance on international assessments of educational achievement: Findings from the condition of education 2006* (No. NCES 2006-073). Washington, DC: U.S. Department of Education.

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

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

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing, 10*, 325-337.

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

Martin, E., Pino, G., & De Boeck, P. (2006). IRT Models for Ability-Based Guessing. *Applied Psychological Measurement, 30*(3), 183-203.

McLachlan, G., & Peel, D. (2000). *Finite Mixture Models*. New York: John Wiley & Sons.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equations of state calculations by fast computing machines. *The Journal of Chemical Physics, 21*, 1087–1091.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*(3), 359-381.

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

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*(4), 557-585.

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10*(2), 133-142.

Muthén, L., & Muthén, B. (1998-2007). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

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

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

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London, A*(185), 71-110.

Pelton, T. W. (2002). *The accuracy of unidimensional measurement models in the persence of deviations for the underlying assumptions.* Unpublished Unpublished doctoral dissertation, Brigham Young University.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 271-286). New York: Springer-Verlag.

Redner, R. A., & Walker, H. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review, 26*, 195-239.

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

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation, 72*, 217-232.

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

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213-232.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin of Moscow, 2*, 3-16.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society, Series B, 64*(4), 583-616.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). WINBUGS 1.4. User Manual. Cambridge: MRC Biostatistics Unit.

Tsutakawa, R. K., & Soltys, M. J. (1988). Approximation for Bayesian ability estimation. *Journal of Educational Statistics, 13*, 117-130.

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

von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York, NY: Springer Science.

von Davier, M., & Rost, J. (2007). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26): Elsevier B. V.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response Theory: An analog for the 3-PL useful in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.),

Computerized adaptive testing: Theory and practice (pp. 245-270). Boston, MA: Kluwer-Nijhoff.

Wilks, W. R., Richardson, S., & Spiegelhalter, D. (Eds.). (1996). *Markov Chain Monte Carlo in practice*: Chapman & Hall/CRC.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*(2), 95-114.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment, 10*(1), 1-17.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19-38.

Yamamoto, K. (1987). *A model that combines IRT and latent class models.* Unpublished doctoral dissertation, University of Illinois Urbana-Champaign.

Yamamoto, K. (1989). *Hybrid model of IRT and latent class models* (No. RR-89-41). Princeton, NJ: Educational Testing Service.

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (No. RR-95-16). Princeton, NJ: Educational Testing Service.

Yang, X. (2007). Methods of identifying individual guessers from item response data. *Educational And Psychological Measurement, 67*(5), 745-764.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. (2003). BILOG-MG 3: Item analysis and test scoring with binary logistic models [Computer software]. Chicago, IL: Scientific Software.