8-1-2017

A Comparison of Two MCMC Algorithms for Estimating the 2PL IRT Models

Meng-I Chang
Southern Illinois University Carbondale, mengi@siu.edu

A COMPARISON OF TWO MCMC ALGORITHMS FOR ESTIMATING THE 2PL IRT MODELS

by

Meng-I Chang

B.S., Chung Yuan Christian University, Taiwan, 2001
M.A., Southern Illinois University Carbondale, 2008

A Dissertation
Submitted in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

Department of Counseling, Quantitative Methods, and Special Education
in the Graduate School
Southern Illinois University Carbondale
August 2017

DISSERTATION APPROVAL

A COMPARISON OF TWO MCMC ALGORITHMS FOR ESTIMATING THE 2PL IRT MODELS

by

Meng-I Chang

A Dissertation Submitted in Partial
Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy
in the field of Counseling, Quantitative Methods, and Special Education

Approved by:
Yanyan Sheng, Ph.D., Chair
Todd Headrick, Ph.D.
Rhonda Kowalchuk, Ph.D.
David Olive, Ph.D.
John Reeve, Ph.D.

Graduate School
Southern Illinois University Carbondale
June 2, 2017

AN ABSTRACT OF THE DISSERTATION OF

MENG-I CHANG, for the Doctor of Philosophy degree in Counseling, Quantitative Methods, and Special Education, presented on June 2, 2017 at Southern Illinois University Carbondale.

TITLE: A COMPARISON OF TWO MCMC ALGORITHMS FOR ESTIMATING THE 2PL IRT MODELS

MAJOR PROFESSOR: Dr. Yanyan Sheng

The fully Bayesian estimation via the use of Markov chain Monte Carlo (MCMC) techniques has become popular for estimating item response theory (IRT) models. The current development of MCMC includes two major algorithms: Gibbs sampling and the No-U-Turn sampler (NUTS). While the former has been used with fitting various IRT models, the latter is relatively new, calling for the research to compare it with other algorithms. The purpose of the present study is to evaluate the performances of these two emerging MCMC algorithms in estimating two two-parameter logistic (2PL) IRT models, namely, the 2PL unidimensional model and the 2PL multi-unidimensional model under various test situations. Through investigating the accuracy and bias in estimating the model parameters given different test lengths, sample sizes, prior specifications, and/or correlations for these models, the key motivation is to provide researchers and practitioners with general guidelines when it comes to estimating a UIRT model and a multi-unidimensional IRT model. The results from the present study suggest that NUTS is equally effective as Gibbs sampling at parameter estimation under most conditions for the 2PL IRT models. Findings also shed light on the use of the two MCMC algorithms with more complex IRT models.

ACKNOWLEDGEMENTS

There are many individuals whom I would like to thank for helping me in completion of this dissertation. First, I would like to express my sincere gratitude and thanks to my advisor, Dr. Yanyan Sheng. Without her encouragement, guidance, and patience throughout my entire doctoral study, this dissertation would never have taken shape. She always made time for me no matter how busy she was when I needed her help. I really appreciate her for unsparingly imparting her knowledge and expertise in this dissertation. I also would like to thank my other committee members, Dr. David Olive, Dr. John Reeve, Dr. Rhonda Kowalchuk, and Dr. Todd Headrick for their insightful comments and valuable suggestions, which inspired me to widen my dissertation from various perspectives.

My appreciation would also goes to my colleagues in the department, present and past, who provided a friendly and supportive environment to work and learn. Also, I would like to thank my friend, Tzu-Chun Kuo who has provided tremendous encouragement and help during my doctoral life.

Last but not the least, I would like to express my gratitude to my parents, brothers, and sister-in-law for their unconditional love and support even though they are 8000 miles away. Thank them for believing in me and helping me achieve my dreams.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Educational and psychological measurement is a field of interest for many researchers to construct objective measurement of individuals' skills, knowledge, and abilities. Classical test theory (CTT, Traub, 1997) is a psychometric theory that can be used to develop and further validate such measurement. The goal of CTT is to understand the characteristics of developed instruments, investigate the performance of individual items, and possibly improve the test's reliability and validity. Although CTT has been broadly used in measurement for decades, it has its shortcomings. The major ones include (a) the person ability (e.g., observed scores) depends on items selected, and (b) the item characteristic (e.g., item difficulty or item discrimination) depends on groups selected (Hambleton & Swaminathan, 1985). Due to these, applications such as test equating and construction (Skaggs & Lissitz, 1986), computerized adaptive testing (CAT; Linden & Glas, 2000), differential item functioning (DIF; Holland & Wainer, 1993), linking and building item banks would be difficult to perform in CTT (Fan, 1998; Güler, Uyanık, & Teker, 2014). Item response theory (IRT; Lord, 1980), an extension of CTT, provides a solution. Instead of focusing on information at the test level as CTT does, IRT mainly postulates the probabilistic relationship between a person's latent trait and the test at the item level. Latent traits are a specific type of constructs that refer to unobservable or unmeasurable objects. They include entities such as attitudes, preferences, and disposition and various underlying processes that educators are interested in measuring such as ability, aptitude, expertise, and intelligence. For example, we can use achievement tests to measure individuals' constructs such as memory. Because of IRT's advantages over CTT, it has gained increased popularity in large-scale educational and psychological testing settings (e.g., Baker & Kim, 2004; De Ayala, 2009; Hambleton & Jones, 1993).

Based on the number of latent constructs being measured, IRT models can be conceived as being unidimensional or multidimensional. Unidimensional IRT (UIRT) models are utilized in circumstances when all test items measure one single latent trait. Dichotomous UIRT models (e.g., Birnbaum, 1969; Lord, 1980; Lord & Novick, 1968; Rasch, 1960) can be applied to cognitive/achievement tests where two response categories such as correct/incorrect or true/false responses are used. Various such models have been developed in the literature, including the conventional one-, two-, and three-parameter models. The one-parameter model (Rasch, 1960; Wright & Stone, 1979) is the simplest IRT model because it only contains the difficulty parameter (i.e., the ability required for individuals to have a probability of 50% to respond to the item correctly). The two-parameter model (Lord, 1952) extends the one-parameter model by adding the discrimination parameter, which is proportional to the slope at the point of the difficulty level. Items with steeper slopes are more useful for separating individuals with different ability levels than are items with less steep slopes. The three-parameter model extends the two-parameter model by adding the pseudo-guessing parameter, which is the probability of individuals with low ability answering the item correctly. With a logit or a probit link, these dichotomous UIRT models can be defined in either logistic or normal ogive forms. In the literature, such models are equivalent (Edelen & Reeve, 2007) in providing similar item characteristic curves (ICCs), which specify that as the level of latent trait increases, the probability of a correct response to an item increases (Hambleton, Swaminathan, & Rogers, 1991). Dichotomous UIRT models have been broadly studied in the literature (e.g. Kang & Cohen, 2007; Rizopoulos, 2006) and are the focus of this dissertation.

UIRT models have two major assumptions: unidimensionality and local independence (Lord, 1980). Unidimensionality states that only one single latent trait is measured by a set of test items. This assumption is related to the assumption of local independence, which means that when the latent trait is held constant, individuals' responses to any pair of items are independent. In other words, the latent trait measured by a test is the only factor that affects the probability of correct responses to individual items. In most situations, all test items are designed to measure one trait and hence it is appropriate to use UIRT models. However, when multiple traits are being measured or the test dimensionality structure is not obvious, using a UIRT model becomes problematic because measurement error inflates and incorrect inferences about an individual's proficiency in a given subject may be made (e.g., Walker & Beretvas, 2000). In such cases, multidimensional IRT (MIRT; Reckase, 1997, 2009) models should be considered.

There are two general forms of MIRT models: compensatory and noncompensatory. For compensatory MIRT models, a lack of one trait dimension can be compensated by an increase of other trait dimensions (e.g., Ackerman, Gierl, & Walker, 2003; Reckase, 1985). However, for noncompensatory MIRT models, a lack of one cannot be offset by an increase of others (e.g., Sympson, 1978; Whitely, 1980). Due to the estimation complexity in noncompensatory MIRT models, most research has focused on compensatory MIRT models (De Ayala, 1992). By using the compensatory form, the multidimensional one-, two-, and three-parameter models can be extended from unidimensional models (De Ayala, 1992; DeMars, 2010). A special case of MIRT models is known as the multi-unidimensional IRT model when the overall test is multidimensional but each item measures only one latent trait (Sheng & Wikle, 2007).

To date, many estimation techniques have been developed for various IRT models with early focus being on using the joint maximum likelihood (JML; Birnbaum, 1969). The JML estimation begins with the joint probability (likelihood) of the item response vector given the person parameters. This procedure treats both item and person parameters as unknown and simultaneously estimates them by maximizing the joint likelihood. The JML, however, tends to result in inconsistent and biased estimates (Andersen, 1970; Gruijter, 1990; Ghosh, 1995; Neyman & Scott, 1948).

The marginal maximum likelihood (MML; Bock & Aitkin, 1981) method based on the expectation maximization (EM) algorithm was developed in the early 1980's to overcome the problems resulting from using the JML estimation. It treats persons as random effects and derives a marginal probability of observing the item response vector by integrating the person effects out of the joint likelihood in order to separate item parameters from person parameters. Therefore, in MML, two steps are taken where item parameters are first estimated using the EM algorithm after integrating out person parameters, and then person parameters can be subsequently estimated by fixing the estimated item parameters as known. Given that both JML and MML produce estimators related to the maximum likelihood, they may result in either infinite or impossible parameter estimates in circumstances where unusual response patterns are observed (e.g., perfect or zero scores).

With the help of modern computer techniques, the estimation methods of IRT models have gradually shifted to the fully Bayesian estimation, which can simultaneously obtain posterior estimates for both item and person parameters. The fully Bayesian estimation via the use of Markov chain Monte Carlo (MCMC; Hastings, 1970) simulation techniques has demonstrated its advantages over traditional maximum likelihood for IRT models (e.g., Kim, 2007; Mislevy, 1986; Swaminathan & Gifford, 1983). Unlike JML and MML, the Bayesian method can avoid unreasonable parameter estimates occurring. In addition, the Bayesian approach controls the parameters within a reasonable range via specifying appropriate prior distributions. Further, the fully Bayesian estimation, with the use of MCMC methods, is highly flexible and has demonstrated its practical usefulness in all aspects of Bayesian inferences, such as parameter estimation or model comparisons.

As the name indicates, MCMC combines Monte Carlo with Markov chain. Monte Carlo is a computational simulation technique with the name coming from the Monte Carlo casino in Monaco. A Markov chain is a stochastic process that satisfies the Markov property if one can make predictions for the future process based solely on its present value. In other words, what happens next in the chain depends only on the current state of the system and not on how it reached the current state. An important feature of a Markov chain is its stationary distribution. The stationary state allows one to define the probability for every state of a system at a random time. Therefore, MCMC methods are a class of algorithms that can be used to simulate samples from a probability distribution via constructing a Markov chain that has the desired distribution (i.e., the posterior distribution) as its stationary distribution.

Common MCMC algorithms include Gibbs sampling (Geman & Geman, 1984) and Metropolis-Hastings (MH; Hastings, 1970; Metropolis & Ulam, 1949). These methods engage in random walk behaviors where at each step, the direction of the proposed move is random. If the relative probability of the proposed position is less than that of the current position, the acceptance of the proposed move is by chance. Due to the randomness, if the process were started over again, then the movement would certainly be different. Gibbs sampling is the simplest MCMC algorithm that requires the marginal distribution of each parameter conditional on the values of all the others to be in closed form. The algorithm works by drawing random samples of each parameter from its full conditional distribution based on the previously generated values of all the other parameters. Then, the joint posterior distribution can be eventually obtained through an adequate number of iterations. If the full conditional distribution is not in closed form or is difficult to simulate, one has to use a more general MH algorithm, which chooses a proposal or candidate distribution by the current value of the parameters. Then, a proposal value is generated from the proposal distribution and accepted in the Markov chain with a certain amount of probability. Although the MH method can be applied in many situations, finding an appropriate proposal distribution for each parameter could sometimes be inefficient in the Markov chain. In addition, both Gibbs sampling and MH utilizing random walks have the general problem of possibly requiring too much time to reach convergence to the target distribution for complicated models with many parameters. These methods tend to explore the parameter space via inefficient random walks (Neal, 1992).

Other MCMC methods such as Hamiltonian Monte Carlo (HMC; Duane, Kennedy, Pendleton, & Roweth, 1987) and No-U-Turn Sampler (NUTS; Hoffman & Gelman, 2014) have been developed to avoid the random walk behavior that Gibbs sampling or MH exhibits by introducing an auxiliary momentum vector and implementing Hamiltonian dynamics so the potential energy function is the target density. HMC generates a proposal in a way similar to rolling a small marble on a hilly surface (the posterior distribution). The marble gains kinetic energy when it falls down the hill and earns potential energy when it climbs back up the hill. The proposed point is then accepted or rejected according to the Metropolis rule. HMC obtains a sequence of random samples from a probability distribution for which direct sampling is difficult. This sequence can be used to approximate the distribution (i.e., to generate a histogram), or to compute an integral (such as an expected value). NUTS further improves HMC by eliminating the need to manually set the number of steps in HMC at each iteration. The algorithm gets its name “no-U-turn” sampler because it prevents inefficiencies that would arise from letting the trajectories make a U-turn. NUTS generalizes the notion of the U-turn to high dimensional parameter spaces and estimates when to stop the trajectories before they make a U-turn back toward the starting point. Gibbs sampling and NUTS can be implemented in software such as JAGS (Plummer, 2003) and Stan (Stan Development Team, 2016), respectively.

1.1 Statement of the Problem

In the IRT literature, many studies have been conducted on the development and application of Bayesian IRT models using Gibbs sampling or MH (e.g., Albert, 1992; Albert & Chib, 1993; Béguin & Glas, 2001; Patz & Junker, 1999a, 1999b; Sheng & Wikle, 2007, 2008, 2009) as well as using NUTS (Caughey & Warshaw, 2014; Zhu, Robinson, & Torenvlied, 2014). Also, studies comparing fully Bayesian and maximum likelihood estimation (MLE) have found that Gibbs sampling performs better than MH (e.g., Sahu, 2002) with the use of data augmentation (Tanner & Wong, 1987) and MLE (e.g., Albert & Chib, 1993) with the small sample size situation. Recently, Grant, Furr, Carpenter, and Gelman (2016) tried to fit the one-parameter IRT model (Rasch, 1960) using both Gibbs sampling and NUTS. Although their results showed that NUTS performed better than Gibbs sampling, their study only focused on the computation speed and scalability. To date, no research has actually investigated the comparison of Gibbs sampling and NUTS in estimating the dichotomous UIRT and multi-unidimensional IRT models. Hence, given the increased popularity of fully Bayesian estimation using Gibbs sampling and NUTS, and the ease in implementing them via two computer programs, JAGS and Stan, it is important and necessary to investigate how these two types of MCMC algorithms perform in estimating item and person ability parameters in such models especially when different sample size, test length, prior specification, and/or intertrait correlation conditions for these IRT models are considered.

1.2 Purpose of the Study

The purpose of the study is to evaluate the performances of two emerging MCMC algorithms, Gibbs sampling and NUTS, in estimating the two-parameter logistic (2PL) UIRT model and the 2PL multi-unidimensional IRT model under various test situations. The parameter estimates were obtained using these two algorithms, which can be implemented in two computer programs, JAGS and Stan, respectively. The key motivation for this investigation is to provide researchers and practitioners with general guidelines when it comes to estimating a UIRT model and a multi-unidimensional IRT model using Gibbs sampling and NUTS. Moreover, the accuracy and bias in estimating the model parameters were investigated given different test lengths, sample sizes, prior specifications, or correlations for these models.

1.3 Research Questions

The general research question is to compare two types of MCMC algorithms, i.e., Gibbs sampling where random walk is utilized with NUTS where random walk behaviors are avoided for the 2PL IRT models. Each algorithm was implemented to the 2PL UIRT and multi-unidimensional IRT models. The specific research questions related to the performance of the model and parameter estimations are as follows

1. How does Gibbs sampling compare with NUTS in estimating the 2PL UIRT model under various test conditions where sample sizes, test lengths, and prior specifications differ?
2. How does Gibbs sampling compare with NUTS in estimating the 2PL multi-unidimensional IRT model under various test conditions where sample sizes, test lengths, and intertrait correlations differ?

1.4 Definition of Terms

For the purpose of this dissertation, some important terms are defined as follows:

- Item response theory (IRT) − Item response theory, also known as the latent trait theory, is the theory used in educational and psychological measurement (e.g., achievement tests, rating scales, and inventories) that investigates a mathematical relationship between individuals' abilities (or other mental traits) and item responses.
- Unidimensional IRT (UIRT) − UIRT assumes each of the individual trait level varies continuously along a single dimension. A person's response to a specific item is determined by a single unified latent trait.
- Multidimensional IRT (MIRT) − MIRT assumes multiple traits are measured by each item.
- Multi-unidimensional IRT − It is a special case of MIRT. It assumes that an overall test measures multiple latent traits, with each subtest measuring one of them. This implies that the overall test is multidimensional while each subtest is unidimensional. The latent traits can be correlated.
- Dichotomous IRT models − A dichotomous IRT model is used when a test involves items with two response categories (e.g., true/false items).
- Fully Bayesian− It is a branch of mathematical probability theory that allows one to model uncertainty about the world and outcomes of interest by combining common-sense (prior) knowledge and observational evidence (likelihood).
- Markov chain Monte Carlo (MCMC) − MCMC methods are a class of algorithms for generating samples from a probability distribution via constructing a Markov chain that has the desired distribution as its stationary distribution. MCMC methods are used in data modeling for Bayesian inference and numerical integration.

- Random walk Monte Carlo methods – A MCMC algorithm can be a random walk that uses either an acceptance or rejection rule to converge to the target distribution. Algorithms such as Gibbs sampling and Metropolis-Hastings algorithm are considered as random walk Monte Carlo methods.
- Gibbs sampling − This is one of the simplest MCMC algorithms. Gibbs sampling is applicable when the joint posterior distribution is not known explicitly, but the conditional posterior distribution of each parameter is known. The idea of a Gibbs sampler is to obtain the joint posterior distribution by iteratively generating a random sample from the full conditional distribution for each parameter.
- Metropolis-Hastings (MH) − This is more general than Gibbs sampling and used when any of the conditional posterior distributions do not have an obtainable closed form. The idea of MH is to generate a proposed value from a proposal distribution. Then the proposed value is accepted as the next value in the Markov chain with a certain probability.
- Hamiltonian Monte Carlo (HMC) −HMC is a MCMC algorithm that uses an auxiliary momentum vector and implements Hamiltonian dynamics so the potential energy function is the target density.
- No-U-Turn Sampler (NUTS) – NUTS is one of the MCMC algorithms that build a set of likely candidate points that span a wide swath of the target distribution, stopping automatically when it starts to double back and retrace its steps.
- JAGS – JAGS stands for just another Gibbs sampler and is a program for analysis of Bayesian models with MCMC simulation using Gibbs sampling.
- Stan − It is a program that performs Bayesian inference using NUTS.

1.5 Significance of the Research

The significance of the study lies in the comparison between two MCMC algorithms for the 2PL dichotomous UIRT and multi-unidimensional IRT models under various test conditions. The results of the study provide researchers and practitioners with a set of guidelines on using Gibbs sampling and NUTS in estimating the two IRT models. Understanding the performance of the two algorithms for IRT models in various test situations would further provide guidance to future research when it comes to parameter estimation for IRT models using the algorithms under investigation. Findings from this study also provide empirical evidence on the use of Gibbs sampling or NUTS with more complicated IRT models.

1.6 Delimitation of the Study

The delimitations in this dissertation are described as follows.

1. This dissertation focuses on the two-parameter dichotomous IRT model. More item parameters such as three-parameter IRT models, or polytomous IRT models such as the partial credit and the rating scale models are not considered.
2. This dissertation only compares two MCMC algorithms under the fully Bayesian framework. Other MCMC algorithms such as Metropolis-Hastings or Hastings-within-Gibbs, or other estimation methods such as JML or MML are not considered.
3. When implementing the IRT models, this dissertation focuses only on simulated data not real data because with simulations, the model parameters can be specified, which makes it possible to evaluate the performance of each estimation algorithm.
4. The two-parameter multi-unidimensional IRT model is a special case of the corresponding MIRT model. It is noted that the multi-unidimensional model does not apply to situations where each item measures multiple latent traits.

5. The simulation study for the multi-unidimensional IRT model in this dissertation only considers situations where a test involves two subscales. The algorithms, however, can be applied to situations where more than two dimensions are involved.

1.7 Overview of Subsequent Chapters

The subsequent chapters are organized as follows. Chapter 2 reviews the related literature on the IRT models, estimation procedures, and the algorithms of implementing IRT models under the fully Bayesian framework. Chapter 3 describes the procedures of fitting the models with simulated datasets. Chapter 4 presents the results of simulation studies for the 2PL UIRT and multi-unidimensional IRT models. Finally, Chapter 5 summarizes the conclusion of findings, implication of this study, and the discussion for future research.

CHAPTER 2

LITERATURE REVIEW

The review of literature starts with the basic concept of item response theory. Five main sections are included in this chapter. The first section reviews the unidimensional and multidimensional IRT models, with the multi-unidimensional model as a special case. The second section focuses on the estimation procedures with UIRT models. Section 3 reviews the estimation procedures with MIRT models. Section 4 concentrates on a few major Markov chain Monte Carlo (MCMC) algorithms and programs that can be used to implement these algorithms. The last section reviews prior research estimating unidimensional and multidimensional IRT models using fully Bayesian estimation via MCMC.

2.1 Item Response Theory

Item response theory (IRT; Lord, 1980) is a measurement theory used in educational and psychological assessments (e.g., achievement tests, rating scales, and inventories) that assumes a mathematical relationship between individuals' abilities (or other mental traits) and item responses (Baker & Kim, 2004; Hambleton et al., 1991; Wainer, Bradlow, & Wang, 2007). IRT is constructed on the concept that the probability of a correct response to an item is a mathematical function of both person and item parameters (Hemker, Sijtsma, & Molenaar, 1995). It is generally considered as an improvement over classical test theory (CTT), which has become the norm for test measurement since the 1930s and has been the predominant psychometric method with psychological instruments for most of the last century (Gulliksen, 1987). For tests that can be performed using CTT, IRT generally provides more flexibility and offers additional test information. Some applications, such as computerized adaptive testing (CAT), are enabled by IRT and cannot reasonably be performed using CTT only. Although CTT has been used in educational and psychological measurement for decades, it has its own limitations. For example, for CTT, the item characteristics are sample dependent, and hence can change based on the groups of individuals that are being selected. Another limitation of CTT is that the traits of individuals depend on the items selected. Thus, it is difficult to compare individuals' latent traits if the test forms are not exactly parallel. IRT was developed to tackle the limitations of CTT and offered more information on test scores. IRT differs from CTT in that it has the property of invariance of item and latent trait characteristics, which indicates that the corresponding estimates are not sample or item dependent. In other words, latent trait estimates from different item sets evaluating the same fundamental construct are similar and vary only because of the examinee measurement error. Item estimates from different respondent groups in the same population are similar and vary only because of the sampling error. The comparisons between IRT and CTT have been widely explored in the literature (see e.g., De Ayala, 2009; Hambleton & Jones, 1993; Thissen & Wainer, 2001).

2.1.1 IRT Major Assumptions

There are two main assumptions with conventional IRT models, including unidimensionality and local independence. Unidimensionality states that only one single latent trait θ is measured with a set of test items. In reality, this assumption can be difficult to meet because several cognitive, personality, and test taking factors directly affect individuals' test performance. To overcome this, a "dominant" factor that affects the test performance is required. Once the assumption of unidimensionality is met, local independence is also obtained. To some extent, these two concepts are equivalent (Lord, 1980). Local independence means that when individuals' latent traits are held constant, their responses to any pairs of items are statistically independent. In other words, after taking individuals' abilities into account, no relationship exists between individuals' responses to test items.

Due to these assumptions, IRT models can be generally divided into two categories: unidimensional and multidimensional IRT models. Unidimensional IRT models assume each of the latent traits varies continuously along a single dimension θ, while multidimensional IRT models are used to measure multiple traits (Reckase, 1997, 2009). However, given the greatly increased complexity involved with multidimensional IRT models, the majority of IRT research and applications focuses on unidimensional IRT models. In addition, based on the number of scored responses, IRT models can also be categorized as models for dichotomous outcomes (e.g., true/false; correct/incorrect), and those for polytomous outcomes, where each response has a different score value. A common example of the latter is Likert-type items (e.g., "Rate on a scale of 1 to 5"). Given that this dissertation focuses on dichotomous models, interested readers can refer to Samejima (1969, 1972), Masters (1982), and Muraki (1992) for polytomous models.

2.1.2 Unidimensional IRT (UIRT) Models

Common dichotomous UIRT models are described by the number of item parameters they consist of. The one-parameter model is the simplest UIRT model. The model contains an item difficulty parameter (b_j), which corresponds to the ability required for individuals to respond to the item correctly at a probability of 0.5. The one-parameter logistic (1PL) model, also known as the Rasch model (Rasch, 1960), is defined as the probability of a correct response for person *i* to item *j* ($Y_{ij} = 1$):

$$P(Y_{ij} = 1|\theta_i, b_j) = \frac{\exp(\theta_i - b_j)}{1+\exp(\theta_i - b_j)}, \qquad (2.1.1)$$

where θ_i is the latent variable of person *i* (*i* = 1,…, *N*). θ_i ranges from $-\infty$ to $+\infty$ and follows a standard normal distribution. Given this, the majority of persons (99.7%) have θ values ranging from -3 to 3 (DeMars, 2010). The range of b_j (j = 1,…, K) is from -2 to 2 in practice when θ_i is assumed to be between -3 and 3 (Hambleton & Cook, 1977). Given that b_j denotes item difficulty, the larger its value is, the more difficult this item becomes since it requires individuals' greater ability to attain a 50% correct response. With a probit form, the one-parameter model can be defined as the one-parameter normal ogive model:

$$P(Y_{ij} = 1|\theta_i, b_j) = \Phi\ (\theta_i - b_j), \tag{2.1.2}$$

where $\Phi(\cdot)$ is the standard normal cumulative density function.

The two-parameter model assumes that items can vary in terms of difficulty (b_j) and discrimination (a_j). The two-parameter logistic (Lord & Novick, 1968) and normal ogive models are defined as follows:

$$P(Y_{ij} = 1|\theta_i, a_j, b_j) = \frac{\exp[a_j(\theta_i - b_j)]}{1+\exp[a_j(\theta_i - b_j)]}\text{, and} \tag{2.1.3}$$

$$P(Y_{ij} = 1|\theta_i, a_j, b_j) = \Phi[a_j(\theta_i - b_j)], \tag{2.1.4}$$

where a_j is referred to as the discrimination parameter for item j. The value of a_j can range from $-\infty$ to $+\infty$, but in practice, it ranges from 0 to 2 (DeMars, 2010; Hambleton & Cook, 1977). An item with a negative discrimination parameter suggests that individuals with greater abilities are less likely to answer the item correctly. Hence, such items should be revised or removed.

The three-parameter model is an extension of the two-parameter model by adding a pseudo-guessing parameter c_j for item j. The three-parameter logistic and normal ogive models are described as

$$P(Y_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1+\exp[a_j(\theta_i - b_j)]}, \tag{2.1.5}$$

$$P(Y_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + \left(1 - c_j\right) \Phi[a_j(\theta_i - b_j)]. \tag{2.1.6}$$

If a five-option multiple choice item is used, c_j would be approximately 0.2, which is the chance that an individual with an extremely low latent trait could answer this item correctly. For example, in multiple-choice aptitude tests, even the least competent people can score by guessing (Drasgow & Schmitt, 2002). For $c_j > 0$, the item difficulty is not the trait level at which the probability that an individual answers correctly is 0.5. Instead, the inflation point (1+c_j)/2 is shifted by the lower asymptote.

In summary, the three-parameter model gets its name because it contains three item parameters, including the difficulty (b_j), discrimination (a_j), and guessing (c_j) parameters. The two-parameter model assumes that items can differ in terms of difficulty (b_j) and discrimination (a_j) with no guessing. The one-parameter model assumes that all items have comparable discriminations and that guessing is a part of the ability, and hence items can be described by a single parameter (b_j).

In addition to the three conventional IRT models, there is a hypothetically four-parameter model (Barton & Lord, 1981), which adds an upper asymptote, represented by d_j. The upper asymptote d_j allows high-ability students to miss an easy item without their ability being drastically underestimated (Barton & Lord, 1981). Therefore, $1-c_j$ in the three-parameter model is replaced by $d_j - c_j$. This model, however, is rarely used. Note that the alphabetical order of the item parameters does not necessarily suggest their practical or psychometric importance. The difficulty parameter (b_j) is clearly the most important because it is included in all four models. The one-parameter model only has b_j, the two-parameter model has b_j and a_j, the three-parameter model adds c_j, and the four-parameter model adds d_j.

The three-parameter model is equivalent to the two-parameter model with c_j = 0, which is appropriate for testing items where guessing the correct answer is highly unlikely, such as fill-in-the-blank questions (“What is the square root of 121?”), or where the concept of guessing does not apply, such as personality, attitude, or interest items (e.g., “Do you like Broadway musicals? Yes/No”).

2.1.3 Multidimensional IRT (MIRT) Models

When multiple latent traits are being measured or the test dimensionality structure is not obvious, it could be problematic to fit the data with a UIRT model since measurement error increases and incorrect inferences about an individual’s proficiency may be made (Walker & Beretvas, 2000). In such cases, multidimensional IRT (MIRT, Reckase, 1997, 2009) models should be used for dealing with this type of complicacy in educational and psychological measurement.

MIRT models have been developed to explain how test items interact with an individual when characteristics of an individual are defined using a vector of hypothetical constructs rather than a single unified trait (Reckase, 1997). The two most common MIRT models are compensatory (e.g., Ackerman et al., 2003; Reckase, 1985) and non-compensatory (e.g., Sympson, 1978; Whitely, 1980) MIRT models. In compensatory MIRT models, a lack of one dimension can be compensated by an increase in other trait dimensions. For example, individuals with a higher arithmetic problem-solving ability might be able to use it to compensate for their lower algebraic symbol manipulation ability in order to correctly respond to a mathematical problem. In contrast, in non-compensatory multidimensional IRT models, a lack of one trait dimension usually cannot be compensated by an increase of others. For example, an individual with a very low reading proficiency attempts solving a math problem. Even with extremely high mathematical skills, that individual will still be unable to solve the math problem described in words.

Due to the difficulties in estimation, more studies have focused on compensatory MIRT models rather than on non-compensatory models (De Ayala, 1992; Knol & Berger, 1991). For example, in an *m*-dimensional test, the compensatory two-parameter logistic MIRT model can be defined as (Reckase, 1985)

$$P(y_{ij}=1|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, \gamma_j) = \text{logit}(\sum_{v=1}^{m} a_{vj}\theta_{vi} - \gamma_j) = \frac{1}{1+\exp[-(\sum_{v=1}^{m} a_{vj}\theta_{vi} - \gamma_j)]}\ , \qquad (2.1.7)$$

where $P(y_{ij}=1|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, \gamma_j)$ is the probability of a correct response of person *i* for item *j*, $\boldsymbol{\theta}_i$= $(\theta_{i1},\ldots, \theta_{im})'$ is an ability vector of person *i* for each of the *m* dimensions, $\boldsymbol{\alpha_j}$ is a vector of discrimination parameters where $\boldsymbol{\alpha}_j = (\alpha_{1j},\ldots, \alpha_{mj})'$, and γ_j is a scalar parameter representing the location in the latent space where the item is maximally informative. When the link function is probit (Φ) rather than logit, the model is called the compensatory two-parameter normal ogive MIRT model. The compensatory two-parameter logistic or normal ogive MIRT models are an extension of the two-parameter logistic or normal ogive UIRT models. Similarly, the compensatory three-parameter logistic (normal ogive) MIRT models can also be extended from the three-parameter logistic (normal ogive) UIRT models (see De Ayala, 1992, for detailed descriptions and equations).

2.1.4 Multi-unidimensional IRT Model

The multi-unidimensional IRT model (Sheng & Wikle, 2007), also known as the between-item MIRT model, can be considered as a special case of MIRT models. For the multi-unidimensional IRT model, items measure only one of the multiple latent abilities, which are commonly in the form of an overall test containing multiple unidimensional subsets or domains (e.g., de la Torre & Patz, 2005; Oshima, Raju, & Flowers, 1997; Sheng & Wikle, 2007; Wang, Wilson, & Adams, 1997). For the multi-unidimensional IRT model, the vector of discrimination parameters in the MIRT model as defined in (2.1.7) is simplified to $\boldsymbol{\alpha}_j$ = (0,…, 0, $\alpha_{vj}, 0, \dots, 0)'$. Specifically, suppose a *K*-item test containing *m* subtests, each having k_v multiple-choice items that measure one trait dimension. With a logit link, the probability of person *i* obtaining a correct response for item *j* of the *v*th subtest can be defined as follows (Lee, 1995):

$$P(y_{vij} = 1 \big| \theta_{vi}, \alpha_{vj}, \gamma_{vj}\big) = \text{logit}\,(\alpha_{vj}\theta_{vi} - \gamma_{vj}) = \frac{1}{1+\exp[-(\alpha_{vj}\theta_{vi}-\gamma_{vj})]}\,, \qquad (2.1.8)$$

where α_{vj} and θ_{vi} are scalar parameters representing the item discrimination and the examinee ability in the *v*th ability dimension, and γ_{vj} is a scalar parameter indicating the location in that dimension where the item provides maximum information. With a probit link, the two-parameter normal ogive multi-unidimensional IRT model can be defined as

$$P(y_{vij} = 1 \big| \theta_{vi}, \alpha_{vj}, \gamma_{vj}\big) = \Phi\,(\alpha_{vj}\theta_{vi} - \gamma_{vj}) = \int_{-\infty}^{\alpha_{vj}\theta_{vi}-\gamma_{vj}} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt. \qquad (2.1.9)$$

2.2 Parameter Estimation of UIRT Models

Accurate recovery of model parameters from response data is a central problem in the IRT models. In fact, successful applications of IRT highly rely on finding appropriate procedures for estimating the model parameters (Hambleton et al., 1991). Numerous estimation techniques have been developed for various IRT models in the past decades. The early method focused on using joint maximum likelihood (JML) and conditional maximum likelihood (CML) estimations. The problem with estimating the item parameters using JML was its tendency to obtain inconsistent estimators (Andersen, 1973). Compared with JML, CML produced more consistent and efficient parameter estimates by removing the trait level parameters from the likelihood equations (Si & Schumacker, 2004). However, when applying the CML estimation to the Rasch model, the parameter estimates are inconsistent due to the loss of item information from the marginal distribution (Andersen, 1973). Bock and Aitkin (1981) presented an algorithm based on the expectation maximization (EM) and since then, the standard approach has been the marginal maximum likelihood (MML) estimation. Modern computer technologies also helped the development of parameter estimation (see Zhao & Hambleton, 2009, for a comparison of the current computer software for IRT analysis) and made it possible to move to the fully Bayesian estimation (e.g., Chib & Greenberg, 1995). Lord (1980) and Baker and Kim (2004) provided a comprehensive review of the methods for parameter estimation with UIRT models. Three main estimation methods, including the JML, MML, and Bayesian estimation are reviewed as follows.

2.2.1 Joint Maximum Likelihood (JML)

The joint maximum likelihood (JML) method relies on the assumption of local independence that individuals' traits are independent of one another and item responses of an individual are independent given the individual's trait θ_i. Therefore, the joint probability (likelihood) of the person parameter θ_i given $\boldsymbol{y}_i$ is

$$L(\theta_i|\boldsymbol{y}_i, \boldsymbol{\xi}) = P(\boldsymbol{y}_i|\theta_i, \boldsymbol{\xi}) = \prod_{j=1}^{k} P(\,\boldsymbol{y}_{ij}|\theta_i,\boldsymbol{\xi}_j), \qquad (2.2.1)$$

where $\boldsymbol{\xi}_j$ is the vector of all item parameters for item j in the IRT model. For example, for the unidimensional two-parameter logistic (2PL) model, $\boldsymbol{\xi}_j = (a_j, b_j)'$ and the likelihood for θ_i is

$$L(\theta_i|\,\boldsymbol{y}_i, \boldsymbol{\xi}) = \frac{\exp\{\theta_i \sum_j y_{ij}a_j - \sum_j y_{ij}a_j b_j\}}{\prod_j\ (1+\exp\{a_j(\theta_i - b_j)\})}. \qquad (2.2.2)$$

The JML method maximizes the joint likelihood function in equation (2.2.1) via simultaneously estimating both item and person parameters. The method treats both item and individual parameters as unknown so the model is unidentified. Therefore, there is no unique solution to find the maximization. To overcome this, constraints have to be placed on the parameters of the model in order to ensure the existence of a solution. Even with constraints, the problem is that the maximization equation cannot be solved analytically unless a numerical method is used. Another problem with the JML method is that the estimations could be inconsistent (Andersen, 1970; Ghosh, 1995; Neyman & Scott, 1948). This is because a limited number of item parameters are estimated in the presence of many person parameters. In that case, regardless of how many individuals are included in the data, the estimation of the item parameters may still be biased (Gruijter, 1990). The JML method is implemented in the LOGIST (Wingersky, 1992) software for one-, two-, and three-parameter IRT models.

2.2.2 Marginal Maximum Likelihood (MML)

The marginal maximum likelihood (MML) method takes a different approach to eliminate the problems encountered in the JML method by treating individuals as random effects and separating person parameter estimation from item parameter estimation via estimating item parameters first. In MML, it is assumed that person parameters θ_i are random effects sampled from a large continuous distribution, denoted $F(\theta)$. The marginal probability of observing the item response vector $\boldsymbol{y}_\mathbf{i}$ is derived by integrating the random person effects out of the joint likelihood defined in (2.2.1), i.e.,

$$P(\boldsymbol{y}_\text{i}|\,\boldsymbol{\xi}) = \int_{\theta_i} L(\theta_i|\boldsymbol{y}_i, \boldsymbol{\xi})dF(\theta_i). \tag{2.2.3}$$

Taking the product of the probabilities in (2.2.3) over individuals i defines the marginal likelihood of the item parameter vector $\boldsymbol{\xi}$:

$$L(\boldsymbol{\xi}|\mathbf{y}) = \prod_i P(\boldsymbol{y}_i|\,\boldsymbol{\xi}). \tag{2.2.4}$$

The MML estimates for the item parameter $\boldsymbol{\xi}$ can be acquired by maximizing the marginal likelihood in (2.2.4) using the EM algorithm. Then, the person parameters θ_i can be obtained using the item parameter estimates. Like the JML method, constraints are needed to identify the model. The constraints can either be placed on the mean and standard deviation of the propensity distribution *F* or on the item parameters. Typically, the distribution *F* is assumed to be the standard normal distribution, but the normal distribution does not necessarily work for all situations. Therefore, it becomes difficult to specify the distribution *F* (Johnson, 2007). In addition, both JML and MML methods encounter the problems that they may result in infinite or impossible parameter estimates. The MML method is directly implemented in BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) for the one-, two-, and three-parameter logistic IRT.

2.2.3 Bayesian Estimation

In the IRT literature, the Bayesian estimation includes a marginal Bayes and a fully Bayesian method. Generally speaking, in the Bayesian approach, model parameters are considered random variables and have prior distributions that reflect the uncertainty about the true values of the parameters before observing the data. The item response models discussed for the observed data are referred to as likelihood models and are the part of the model that presents the density of the data conditional on the unknown model parameters. Therefore, two modeling stages can be recognized: (1) the specification of a prior and (2) the specification of a likelihood model. After observing the data, the prior information is combined with the information from the data and a posterior distribution is constructed. Bayesian inferences are made conditional on the data, and inferences about parameters can be made directly from their posterior densities.

The marginal Bayes estimation uses similar ways for estimating IRT models as the MML method, but the difference is that it places a prior distribution for each parameter in the model. For example, in the three-parameter logistic (3PL) model, the discrimination parameter can be specified to follow a log normal distribution, the difficulty parameter is specified to follow a normal distribution, and the guessing parameter can follow a beta distribution. Then, with information from data (the likelihoods), posterior estimates of item parameters (usually in the form of the mode of the posterior distribution) can be obtained using these priors.

On the other hand, the fully Bayesian estimation can simultaneously obtain posterior estimates for both item and person parameters, and can find the mean of the posterior distribution. Wollack, Bolt, Cohen, and Lee (2002) suggested that fully Bayesian estimation provides a solution when the MML method is not applicable. For decades during the early stage of the development of IRT, the fully Bayesian estimation was not computationally practical for models with a very large number of parameters such as IRT models and therefore, the MML and the marginal Bayes have been the standard estimation methods. Modern computational technology and the development of Markov chain Monte Carlo (MCMC; Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Metropolis & Ulam, 1949) algorithms, however, have made the fully Bayesian estimation applicable to fit different IRT models (e.g., Béguin & Glas, 2001; Bolt & Lall, 2003; Bradlow, Wainer, & Wang, 1999; de la Torre, Stark, & Chernyshenko, 2006; Fox & Glas, 2001; Johnson & Sinharay, 2005; Patz & Junker, 1999a). MCMC methods are a class of algorithms for sampling from a probability distribution (e.g., the posterior distribution) based on constructing a Markov chain that has the desired distribution as its stationary distribution. At each state of the Markov chain, random samples of model parameters are generated from the distribution based on those generated from a previous state. Since early samples may be affected by initial values, they are discarded in the so-called burn-in stage. After the burn-in stage, the quality of the sample becomes approximately stable. Different MCMC algorithms have been developed in the last two decades, and a review of the major ones is made in a later section of this chapter. MCMC methods have been proven useful in practically all aspects of fully Bayesian inference, such as parameter estimation and model comparisons. Albert (1992) was the first to apply the fully Bayesian estimation with IRT by fitting the two-parameter normal ogive (2PNO) IRT model using an MCMC algorithm. Since then, other UIRT models have been developed under the fully Bayesian framework such as two- and three-parameter logistic models (Patz & Junker, 1999a, 1999b) and three-parameter normal ogive models (Sahu, 2002).

Many studies have demonstrated advantages of Bayesian estimation, including the marginal Bayes and the fully Bayesian estimation, over MML and JML methods (e.g., Kim, 2007; Mislevy, 1986; Swaminathan & Gifford, 1983; also see Appendix A for a demonstration of the advantages of fully Bayesian estimation over MML). For example, with the specified prior distribution of item parameters, the Bayesian method avoids the possibility of having unreasonable parameters using MML and JML methods. The specified priors can pull extreme estimates back toward the center of their respective distributions and stop them from assuming unreasonable values. This effect should be more noticeable for small samples and short tests (e.g., Lim & Drasgow, 1990). Even with larger samples and longer tests, however, Bayesian estimation is still superior to JML and MML methods when unusual response patterns occur. For example, individuals may answer all items correctly or incorrectly, or they may answer easy items incorrectly while answering difficult items correctly. Under these circumstances, JML and MML methods would not be able to find an estimate while the Bayesian method will still estimate the parameters within a reasonable range (Baker, 1987; Swaminathan & Gifford, 1983).

The fully Bayesian method also has advantages over the marginal Bayes estimation. Specially, in the marginal Bayes method, person parameters θ_i are treated as random variables and integrated out from the joint likelihood of item and person parameters. However, when the model gets complex, integrating out θ_i could be difficult and as a result, it becomes challenging to implement marginal Bayes. On the other hand, the fully Bayesian approach circumvents the problem of integrating out θ_i since it simultaneously draws samples from the posterior distributions of model parameters.

2.3 Parameter Estimation of MIRT Models

The estimation procedures for MIRT models are relatively complicated for the following reasons (Reckase, 2009). First, as with UIRT, the models contain both person and item parameters and generally, it is difficult to estimate the two sets of parameters independent of each other. Second, compared to UIRT models, MIRT models have more parameters that need to be estimated and hence are more complex. A third reason is that there are indeterminacies in the models such as the location of the origin of the space, the units of measurement for each coordinate axis, and the orientation of the coordinate axes relative to the locations of the persons. All of these issues must be addressed in the construction of an algorithm for estimating the model parameters.

Bock and Aitkin (1981) developed an EM algorithm (Dempster, Laird, & Rubin, 1977) to estimate the parameters of the one-, two-, and three-parameter normal ogive MIRT models for dichotomous items. The algorithm, however, is limited to small testing situations (Baker & Kim, 2004) and it does not work well with a large number of dimensions, either. More estimation methods and computer software have subsequently been developed using the MML technique. For example, Bock, Gibbons, and Muraki (1988) used the MML method and EM algorithm for dichotomous MIRT models and discussed technical problems of its implementation for a number of simulated and real datasets. TESTFACT (Wilson, Wood, & Gibbons, 1991) is a computer program that can be used to implement a nonlinear, exploratory factor analysis for dichotomous test items. The program uses the MML method with an EM algorithm to estimate item parameters in the MIRT model. Then, person parameters are estimated using the Bayesian method by fixing the item parameters.

Trying to use the MML method for MIRT models has been impeded by the fact that the computations involve a numerical integration over the latent ability distribution. With an increased dimensionality, the ability is a multivariate distribution involving more parameters. This makes integration to be computationally demanding such that the applicability of higher dimensional IRT models is impossible in practical settings (Rijmen, 2009). Therefore, more and more researchers have resorted to the fully Bayesian estimation for MIRT models. For example, the algorithm introduced by Albert (1992) for the unidimensional two-parameter normal ogive model was extended to the dichotomous MIRT models (Béguin & Glas, 2001). Other applications of Bayesian estimation of multidimensional dichotomous IRT models can be seen in various studies (e.g., Lee, 1995; Sheng & Headrick, 2012; Sheng & Wikle, 2007, 2008, 2009; Yao & Boughton, 2007; Zheng, 2000).

2.4 MCMC Algorithms

The concept of MCMC methods is to generate samples from a probability distribution via constructing a Markov chain that has the desired distribution as its stationary distribution. As the name shows, MCMC starts with Monte Carlo, a computational simulation technique with a catchy name, i.e., the Monte Carlo casino in Monaco. A Markov chain is a sequence of random variables, $\{X_0, X_1, X_2, \ldots\}$, sampled from the distribution $p(X_{k+1}|X_k)$. Then, each subsequent sample X_{k+1} depends on the current state X_k rather than on previous history $\{X_0, X_1, X_2, \ldots, X_{k-1}\}$. An important concept of a Markov chain is its stationary distribution. The stationary state allows one to define the probability for every state of a system at a random time. Under the fully Bayesian framework, MCMC methods are a class of algorithms that can be used to simulate samples from the posterior distribution and these posterior samples can then be used to summarize the posterior distribution.

Using MCMC-based fully Bayesian methods for estimating complex psychometric models such as IRT models has become popular in recent years. These sampling based methods are more flexible and can provide a more complete picture of the posterior distribution of all parameters in the model than JML or MML does. They can be applied in situations (e.g., small sample size) where the likelihood methods fail or are difficult to implement. The samples produced by the MCMC procedure can also be used for conducting model fit diagnosis, model selection, and model-based prediction.

Common MCMC methods are performed under the notion of random walks, which imply that at each step, the direction of the proposed move is random. If the relative probability of the proposed position is more than that of the current position, then the proposed move is always accepted. If the relative probability of the proposed position, however, is less than that of the current position, the acceptance of the proposed move is by chance. Due to the randomness, if the process were started over again, then the movement would certainly be different. However, regardless of the specific movement, in the long run the relative frequency of visits will be close to the target distribution. The random walk nature of the algorithms can greatly increase the number of iterations required before convergence is reached and/or the number of subsequent iterations that are needed to gather a sample of states from which accurate estimates for the quantities of interest can be obtained. To overcome such inefficiency, other MCMC algorithms such as Hamiltonian Monte Carlo (HMC; Duane et al., 1987) have been developed to reduce random walk behaviors. These algorithms are reviewed as below.

2.4.1 Random Walk MCMC Algorithms

Two fundamental random walk MCMC algorithms, including Gibbs sampling and Metropolis-Hastings (MH) are described as follows. Gibbs sampling, originally introduced by Geman and Geman (1984), is named after the physicist Josiah Willard Gibbs (1839-1903). The process for Gibbs sampling is considered as a type of random walk through the parameter space. The walk begins at some arbitrary point, and at each point in the walk, one of the component parameters is selected and the parameters are cycled through in order (e.g., $\theta_1, \theta_2, \theta_3,$ ....., $\theta_1, \theta_2, \theta_3 \ldots..$). By generating a random value directly from the conditional probability distribution, a new value is selected for that parameter. The process then repeats: select a component parameter and generate a new value for that parameter from its conditional posterior distribution. By cycling through these conditional statements, the joint posterior distribution would be eventually reached. Suppose a multivariate distribution, $p(\boldsymbol{\theta}) = p(\theta_1, \theta_2, \ldots, \theta_p)$ of the random vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is generated and the algorithm of Gibbs sampling can be described as follows:

1. Establish initial values of the parameters, $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)})$.
2. For each iteration, generate a random sample from the distribution of that parameter conditioned on all other parameters, making use of the most recent values and updating the parameter with its new value as soon as it has been sampled. In the *k*th iteration, the *i*th parameter, $\theta_i^{(k)}$ is specified by $p(\theta_i|\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_{i-1}^{(k)}, \theta_{i+1}^{(k-1)}, \ldots, \theta_p^{(k-1)})$, which is the full conditional distribution of θ_i.
3. Repeat step 2 *N* times to get the values $(\boldsymbol{\theta}_i{}^{(0)}, \ldots, \boldsymbol{\theta}_i{}^{(N)})$ for estimating the joint distribution $p(\boldsymbol{\theta}_i)$.

Gibbs sampling is applicable when the joint distribution is not known explicitly or is difficult to sample from directly, but the conditional distribution of each parameter is known and is easy (or at least easier) to sample from. Gibbs sampling, however, does not work well when any of the full conditional distribution is not in closed form. Due to that, the MH (Hastings, 1970; Metropolis & Ulam, 1949) algorithm can be used to estimate parameters. For the MH method, a proposal or candidate distribution is chosen given the current value of the parameter rather than simulating from the full conditional distribution like Gibbs sampling did. The algorithm of MH proceeds as below:

1. Establish initial values of the parameters, $\boldsymbol{\theta}^{(0)}$ = ($\theta_1^{(0)}$, $\theta_2^{(0)}$,…, $\theta_p^{(0)}$).
2. For iterations 1 to *N*, a proposal value, $\boldsymbol{\theta}^{(k)}$, is generated from the proposal distribution, $q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k-1)})$. Then, the proposal value is accepted as the next value in the Markov chain with the probability α = min $\{\frac{p(\boldsymbol{\theta}^{(k)})q(\boldsymbol{\theta}^{(k-1)}|\boldsymbol{\theta}^{(k)})}{p(\boldsymbol{\theta}^{(k-1)})q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k-1)})}$, 1$\}$. If the proposal value is not accepted, the current value would be used as the next value of the Markov chain.
3. Return the values ($\boldsymbol{\theta}^{(0)}$,…, $\boldsymbol{\theta}^{(N)}$) for estimating the joint distribution $p(\boldsymbol{\theta})$.

It is noted that Gibbs sampling could be considered as a special case of the MH algorithm when the probability of accepting the proposal value is always equal to one (Gelman, 2014; Tanner, 1996). In the Metropolis algorithm, a random walk is taken through the parameter space, favoring parameter values that have a relatively high posterior probability. In order to proceed to the next step in the walk, there is a proposed jump from the current position, with the jump sampled randomly from a proposal distribution. The proposed jump could be either accepted or rejected probabilistically, according to the relative densities of the posterior at the proposed position and the current position. If the posterior density is higher at the proposed position than at the current position, the jump is definitely accepted. If the posterior density is lower at the proposed position than the current position, the jump is accepted only with a probability equal to the ratio of the posterior densities. In addition, the step size is a critical tuning factor in the random walk MCMC algorithms. If the average step size is too large, the proposed jump will be rejected almost every cycle and result in high autocorrelation. If the step size is too small, the chain will move very slowly through the parameter space and result in high autocorrelation. Gibbs sampling, where the movement is chosen using a conjugate distribution so the Metropolis-Hastings ratio always accepts, is not necessarily better. Therefore, the effective step size of the Gibbs sampler tends to be small resulting in high autocorrelation (Almond, 2014).

2.4.2 Other MCMC Algorithms

Hamiltonian Monte Carlo (HMC, Duane et al., 1987), also known as a hybrid Monte Carlo, is a MCMC algorithm that tries to avoid the random walk behavior by introducing an auxiliary momentum vector and implementing Hamiltonian dynamics so the potential energy function is the target density. In HMC (Neal, 1992, 2011; Duane et al., 1987), once the proposed jump is established, then the proposal is either accepted or rejected according to the Metropolis decision rule except that the terms involve not only the relative posterior density, but also the momentum at the current and proposed positions. The initial momentum applied at the current position is drawn randomly from a simple probability distribution such as a normal (Gaussian). Denote the momentum as $\emptyset$. Then the Metropolis acceptance probability for HMC is defined as below:

$$p_{accept} = \min(\frac{p(\theta_{proposed}|D)p(\emptyset_{proposed})}{p(\theta_{current}|D)p(\emptyset_{current})}, 1). \tag{2.4.1}$$

In an idealized continuous condition, the sum of potential and kinetic energy [corresponding to $-\log(p(\theta|D))$ and $-\log(p(\emptyset))$] is a constant, and therefore the ratio in (2.4.1) would be one, and the proposal would never be rejected. The end result of HMC is that proposals move across the sample space in larger steps; therefore, they are less correlated and converge to the target distribution more rapidly. HMC uses a proposal distribution that changes depending on the current position and takes a series of steps informed by the first-order gradient information. HMC's performance, however, is highly sensitive to two user-specified parameters: a step size ϵ and a desired number of steps L (Hoffman & Gelman, 2014). The step size regulates the smoothness or jaggedness of the trajectory. The overall duration, steps $(L) *$ step size (ϵ), regulates how far the proposal explores from the current position. It is important to tune this duration since we want the proposal to be close to a mode, without overpassing, and without rolling all the way back to the starting point. In particular, if L is too small, then the algorithm exhibits inefficient random walk behavior, while if L is too large the algorithm wastes computation time. Also, HMC requires the gradient of the log-posterior. It is sometimes impossible to compute the gradient for a complex model, but this requirement can be achieved by using automatic differentiation (Griewank & Walther, 2008). Given the above reasoning, Hoffman and Gelman (2014) introduced the No-U-Turn Sampler (NUTS), an adaptation to HMC that eliminates the need to set a number of steps L. NUTS utilizes a recursive algorithm to construct a set of possible candidate points that crosses a wide strip of the target distribution, stopping automatically when it starts to double back and retrace its steps. Empirically, NUTS performs as well as (and sometimes better than) a well-tuned standard HMC method, without involving user intervention or costly tuning runs (Hoffman & Gelman, 2014).

2.4.3 Implementation of MCMC

One of the primary challenges in implementing MCMC, however, is the availability of accessible software. This issue can be resolved via two emerging computer programs specially developed for implementing any MCMC procedure to a model: JAGS (Plummer, 2003) and Stan (Stan Development Team, 2016).

JAGS, which stands for just another Gibbs sampler, was written in the C++ programming language for Bayesian hierarchical models (Plummer, 2003). JAGS succeeds the pioneering system Bayesian inference using Gibbs Sampling (BUGS, Gilks, Thomas, & Spiegelhalter, 1994), which is implemented in three software packages: WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000; Spiegelhalter, Thomas, Best, & Lunn, 2003), OpenBUGS (Spiegelhalter, Thomas, Best, & Lunn 2010; Thomas, O'Hara, Ligges, & Sturtz, 2006), and JAGS (Plummer, 2003). JAGS can be called from R (R Core Team, 2016) using the R packages rjags (Plummer, 2013) and R2jags (Su & Yajima, 2012). It was written with three aims in mind: 1) to act as a cross-platform engine for the BUGS language, 2) to be flexible, allowing users to write their own functions, distributions and samplers, and 3) to be a platform for experimentation with ideas in Bayesian modeling. With many additional desired features, JAGS has been preferred over BUGS (Plummer, 2003).

Stan, named after the mathematician Stanislaw Ulam (1909-1984), is an open-source C++ program that performs Bayesian inference (Stan Development Team, 2016). Stan uses NUTS (Hoffman & Gelman, 2014), an adaptive variant of HMC (Neal, 2011), which itself is a generalization of the familiar Metropolis algorithm, performing multiple steps per iteration to move more efficiently through the posterior distribution. Gelman, Lee, and Guo (2015) pointed out that compared to BUGS and JAGS, the modeling languages of Stan are more flexible and general. For large data sets or complex models, Stan can provide solutions when JAGS (or BUGS) takes too long or fails. For example, models with matrix parameters (e.g., multilevel models with multiple coefficients that vary by group) are particularly slow in BUGS and JAGS, indicating Gibbs sampling does not work well with covariance matrices (Gelman et al., 2015). Although Stan is not surprisingly fast in such conditions, for problems of moderate size, it runs well enough to be useful. For example, the hierarchical time-series model in Ghitza and Gelman (2014) took several hours to run in Stan but would not have been doable at all in other similar Bayesian software.

2.5 Prior Research Using Fully Bayesian with IRT Models

Computational techniques based on MCMC algorithms have enabled IRT model estimation under the fully Bayesian framework (see e.g., Gilks, Richardson, & Spiegelhalter, 1996 for a review). Some relevant research is reviewed as follows. Albert and Chib (1993) proposed Gibbs sampling for the 2PNO UIRT model and compared the item parameters estimates with those obtained using the maximum likelihood estimation (MLE). Their results showed that Gibbs sampling is preferable to the MLE for small samples and is easier to implement in computer programs. Since then, many studies have been conducted on the development and application of Bayesian UIRT models and MIRT models using random walk MCMC algorithms.

2.5.1 UIRT Models Using Random Walk MCMC Algorithms

Baker (1998) compared Gibbs sampling and MML for a normal ogive IRT model and found that the item parameter estimation was excellent for the largest datasets (50 items and 500 examinees) using Gibbs sampling but for the rest of the test lengths (10, 20, and 30 items) and sample sizes (30, 60, and 120 examinees), the MML performs better than Gibbs sampling in item parameter recovery. Patz and Junker (1999a, 1999b) used an MCMC method called Metropolis-Hastings within Gibbs (Chib & Greenberg, 1995) for the 2PL and 3PL UIRT models and the algorithm performed better than MML in fitting more complex models. Ghosh, Ghosh, Chen, and Agresti (2000) examined noninformative priors for the one-parameter IRT models using Gibbs sampling and found that such priors performed as well as proper priors for item difficulties. Janssen, Tuerlinckx, Meulders and De Boeck (2000) proposed a 2PL hierarchal IRT model using Gibbs sampling and their findings indicated that the recovery was very good for the difficulty parameters but not so good for the discrimination parameters. Fox and Glas (2001) implemented Gibbs sampling to the multilevel 2PNO IRT model and their algorithm worked well in resulting in accurate item parameter estimates. Sahu (2002) compared Gibbs sampling and MH for fitting three-parameter normal ogive (3PNO) IRT models and suggested that Gibbs sampling with the use of data augmentation (Tanner & Wong, 1987) is preferred to MH. Eaves et al. (2005) fitted genetic IRT models using Gibbs sampling and concluded that the algorithm provides a convenient and flexible alternative compared with the MLE for estimating the parameters of IRT models for relatively large data sizes with multi-category items. Sheng (2010) further examined the performance of Gibbs sampling for the 3PNO IRT model with various test-length and sample-size conditions and her findings were that the algorithm was influenced more by the choice of prior specification for the 3PNO model than the 2PNO model. Culpepper (2015) proposed a model for the four-parameter normal ogive (4PNO) IRT model using Gibbs sampling and the results supported the use of less informative uniform priors for the lower and upper asymptotes, and suggested that modest sample sizes (i.e., at least $N = 2500$) are needed to accurately recover all of the 4PNO item parameters.

2.5.2 MIRT Models Using Random Walk MCMC Algorithms

Fully Bayesian estimation using random walk MCMC algorithms has also made parameter estimation possible for MIRT models. Béguin and Glas (2001) used Gibbs sampling for the 3PNO MIRT model and their results showed that Gibbs sampling recovers the true parameter values to a reasonable extent. Hoijtink and Molenaar (1997) examined model parameter estimation and model fit of nonparametric MIRT models using Gibbs sampling and found that Gibbs sampling is an excellent tool if inequality constraints have to be taken into consideration. Bolt and Lall (2003) used MH to evaluate parameter recovery for the multidimensional two-parameter logistic model (M2PL) and the multidimensional latent trait model (MLTM) under various sample sizes, number of items, and correlation between abilities. Their results suggested that parameters of both models can be recovered but is less successful for the MLTM as the correlation between abilities increases. de la Torre and Patz (2005) used the MH algorithm to fit 3PL multi-unidimensional IRT models and the results showed that when taking correlation into account, the multi-unidimensional model resulted in better ability estimates than the unidimensional IRT models. Sheng and Wikle (2007) used Gibbs sampling to fit multi-unidimensional IRT models under the situation when the overall test consists of unidimensional subtests. Their finding indicated that the model provides better results to test situations than the unidimensional IRT model. Then, Sheng and Wikle (2008) proposed MIRT models with a hierarchical structure using Gibbs sampling. The results showed that the proposed models describe the actual data better than the conventional IRT models. Also, Sheng and Wikle (2009) proposed an additive MIRT model using Gibbs sampling and their results showed that the proposed model works well for item parameter estimation if there is no or low correlation between the general and each specific ability. Huo et al. (2015) proposed a hierarchical multi-unidimensional 2PL IRT model using both MH and Gibbs sampling. Their findings were that item parameter could be recovered accurately and the estimated latent trait closely approximated true latent scores.

IRT models have also been developed under the fully Bayesian framework using Gibbs sampling or MH to account for multiple raters (Patz & Junker, 1999b), testlet structures (Bradlow, Wainer & Wang, 1999; Wainer, Bradlow, & Du, 2000; Wang, Bradlow, & Wainer, 2002), latent classes (Hoijtink & Molenaar, 1997), and multidimensional latent abilities (Adams, Wilson, & Wang, 1997; Béguin & Glas, 1998; DeMars, 2005).

2.5.3 IRT Models Using NUTS

To date, there have not been many Bayesian IRT studies conducted using NUTS. Caughey and Warshaw (2014) developed a new group-level hierarchical IRT model using NUTS to estimate dynamic measures of public opinion at the sub-national level and their results showed that this model has considerable advantages over an individual-level IRT model for measuring aggregate public opinion. Copelovitch, Gandrud, and Hallerberg (2015) used a hierarchical Bayesian IRT model implemented in NUTS to develop a new Financial Regulatory Transparency (FRT) Index. The FRT Index is used to measure a country's latent willingness to report minimally credible data about its financial system to international organizations and investors. The results indicated that borrowing costs are less volatile when investors are better able to anticipate instability because they have access to financial regulatory information. Recently, Grant et al., (2015) tried to fit the Rasch (1960) model in both Gibbs sampling and NUTS. Their study, however, only focused on the computation speed and scalability, and the results showed that NUTS performed better than Gibbs sampling as far as these aspects are considered.

In summary, IRT has gained an increasing popularity in large-scale educational and psychological testing situations because of its theoretical advantages over CTT. With current enhanced computational technology and the emergence of MCMC simulation techniques (e.g., Chib & Greenberg, 1995), the methodology for parameter estimation with IRT models has rapidly moved to a fully Bayesian approach. The current development of MCMC focuses on two major algorithms: Gibbs sampling and NUTS, which are implemented in two specialized software packages JAGS and Stan, respectively. Both Gibbs sampling and NUTS show their advantages in efficiently performing Bayesian posterior inference on a large class of complex, high-dimensional models with minimal human intervention and have been applied to IRT models (e.g., Stan Development Team, 2016; Zhu et al., 2014). However, to date, no research has compared the performance of them in fitting more complex IRT models.

CHAPTER 3

METHODOLOGY

This chapter illustrates the methodology that was used to answer the two research questions, which are reiterated in Section 3.1. Specifically, Monte Carlo simulations were carried out to fit the two-parameter logistic (2PL) unidimensional IRT (UIRT) model and 2PL multi-unidimensional IRT model. The details of the simulation studies are provided in Section 3.2 and 3.3.

3.1 Research Questions

The major purpose of this dissertation is to compare the performance of two MCMC algorithms, namely, Gibbs sampling and NUTS, when implementing them to 2PL IRT models. The specific research questions related to the performance of the parameter estimations are as follows:

1. How does Gibbs sampling compare with NUTS in estimating the 2PL UIRT model under various test conditions where sample sizes, test lengths and prior specifications differ?
2. How does Gibbs sampling compare with NUTS in estimating the 2PL multi-unidimensional IRT model under various test conditions where sample sizes, test lengths and intertrait correlations differ?

Two simulation studies were conducted with each addressing one of the two research questions.

3.2 Simulation Study 1

Monte Carlo simulations were conducted to answer research question one via examining the recovery of the model parameters using Gibbs sampling and NUTS under various test conditions.

3.2.1 Model

In simulation study 1, the focus is on the 2PL UIRT model, which is defined as:

$$P(Y_{ij} = 1|\theta_i, a_j, b_j) = \frac{\exp[a_j(\theta_i - b_j)]}{1+\exp[a_j(\theta_i - b_j)]}, \qquad (3.2.1)$$

where Y_{ij} is the probability that the *i*th individual responds to the *j*th item correctly (Y_{ij} = 1) or incorrectly (Y_{ij} = 0), θ_i is the latent ability for subject *i*, a_j is the discrimination parameter, and b_j is the difficulty parameter for item *j*.

In practice, θ_i ranges from -3 to 3. The discrimination parameter is a measure of the differential capability of an item. A high discrimination parameter value suggests an item that has a high ability to differentiate subjects and the usual range for a_j is from 0 to 2. The item difficulty parameter measures the difficulty of answering the item correctly, and in practice, b_j has a range from -2 to 2.

3.2.2 Simulation Procedure

Since sample size and test length play a role in parameter estimation, the general guidelines for stable parameter estimates have been mentioned in the literature (Baker, 1998; Bolt & Lall, 2003). For example, Hulin, Lissak, and Drasgow (1982) used a Monte Carlo study to assess the accuracy of both item and person parameter estimations in IRT. Samples of 200, 500, 1000, and 2000 examinees and tests of 15, 30, and 60 items were generated for the 2PL IRT models. Their results indicated that the minimum sample sizes and test lengths depend on the response model and the purposes of the investigation. With the 2PL model, samples of 500 examinees and tests of 30 items appear adequate for parameter recovery.

Given these, data were generated from the 2PL UIRT model as defined in Equation (3.2.1). Sample size (*N*) was manipulated to be 100, 300, 500, and 1,000 examinees and test length (*K*) was manipulated to be 10, 20, and 40 items. Model parameters were generated such that θ_i is from a normal distribution, θ_i~ N(0, 1), a_j is from a uniform distribution, $a_j \sim U$(0, 2), and b_j is from a uniform distribution, $b_j \sim U(-2, 2)$.

For the MCMC procedures, normal priors were assumed for both θ_i and b_j such that θ_i ~ N(0, 1) and $b_j \sim N(0, 1)$. Three prior specifications were considered for a_j such that

1. a_j was from a lognormal distribution, a_j~ *lognormal*(0, 0.5), which is commonly used in BILOG-MG (Zimowski et al., 2003).
2. a_j was from a truncated normal distribution, a_j~ $N_{(0,\infty)}$ (0, 1), which is another common way to specify the discrimination parameter in the IRT literature (Sahu, 2002; Sheng, 2008; Spiegelhalter et al., 2003).
3. a_j was transformed to α_j such that a_j = exp (α_j), where a standard normal prior was assumed for α_j such that α_j~N(0, 1). With this transformation, any real value exponentiated is positive. Therefore, with this prior specification, Equation (3.2.1) can be reexpressed as

$$\text{logit } (p_{ij}) = \exp\ (\alpha_j)\ (\theta_i - b_j). \qquad (3.2.2)$$

Gibbs sampling and NUTS were implemented for each simulated data set via the use of JAGS and Stan, respectively. In JAGS, the burn-in stage was set to 3000 iterations followed by 4 chains with 5000 iterations. In Stan, the procedure was very similar to JAGS except that Stan uses “warm-up” instead of “burn-in.” Therefore, in Stan, the number of warm-ups was set to 3000 iterations followed by 4 chains with 5000 iterations. For both algorithms, the initial values for the discrimination parameters a_j were set to ones, and those for the difficulty parameters b_j and latent ability parameters θ_i were set to zeros. Convergence was evaluated using the Gelman-Rubin R statistic (Gelman & Rubin, 1992). To obtain the R statistic, multiple Markov chains are generated with spread initial points from the parameter space. Then, the Gelman-Rubin R statistic can be obtained by comparing the variance within and between simulated chains. The procedure is described as follows. Suppose ξ is the parameter of interest. Further, suppose *M* Markov chains were generated, each with a length of L after initial draws are thrown away (burn-in). Denote ξ_{im} as the simulated parameter in the *i*th generation of the *m*th chain. The between sequence variance is defined as B = $\frac{L}{M-1}\sum_{m=1}^{M}(\overline{\xi}_{.m} - \overline{\xi}_{..})^2$, and the within sequence variance is defined as W = $\frac{1}{M}\sum_{m=1}^{M} s_g^2$, where $s_g^2 = \frac{1}{L-1}\sum_{i=1}^{N}(\overline{\xi}_{im} - \overline{\xi}_{i.})^2$. An R statistic is obtained as $\hat{R}$ $=\sqrt{\frac{\widehat{var}(\xi|\boldsymbol{y})}{W}}$, where $\widehat{var}(\xi|\boldsymbol{y})$ = $\frac{L-1}{L}$ W + $\frac{1}{L}$ B. Brooks and Gelman (1998) pointed out that $\hat{R}$ < 1.1 or 1.2 provides evidence that the chain has converged to the posterior distribution. If the R statistic, however, is larger than 1.2 (e.g., $\hat{R}$ = 1.53), it suggests that the Markov chains have not reached stationarity and more iteration runs are needed to improve convergence.

3.2.3 Measures of Estimation Accuracy

For each simulated condition out of the total of sample sizes (4) × test lengths (3) × prior specifications for a_j (3) × algorithms (2) = 72 experimental conditions, 25 replications when *N*=100 and 300, or 10 replications when *N*=500 and 1000 were conducted to avoid erroneous results in estimation due to sampling error. Although Harwell, Stone, Hsu, and Kirisci (1996) suggested a minimum of 25 replications for typical IRT-based Monte Carlo studies, the current study only carried out 10 replications for large data sizes due to the computational expense of the MCMC algorithms for test conditions such as *N*=1000 and *K*=40. The accuracy of item parameter estimates was evaluated using *bias*, the root mean square error (*RMSE*), and the mean absolute error (*MAE*). *Bias* is defined as:

$$bias_{\pi} = \frac{\sum_{j=1}^{n}(\hat{\pi}_j - \pi_j)}{n}, \qquad (3.2.3)$$

where π (e.g., a_j or b_j) is the true value of an item parameter, $\hat{\pi}$ is the estimated value of that parameter in the *k*th replication using either Gibbs sampling or NUTS, and *n* is the total number of replications. If *bias* is close to zero, it suggests that the value of the estimated parameter is close to the true parameter. Also, positive bias suggests that the true parameter is overestimated and a negative bias suggests an underestimation of the true parameter (Dawber, Roger, & Carbonaro, 2009).

The *RMSE* for each item parameter was calculated using the following formula:

$$RMSE_{\pi} = \sqrt{\frac{\sum_{j=1}^{n}(\hat{\pi}_j - \pi_j)^2}{n}}, \qquad (3.2.4)$$

where π, $\hat{\pi}$, and *n* are as defined in Equation (3.2.3).

The *RMSE* measures the average squared discrepancy between a set of estimated and true parameters and can be conceived as the amount of variability around a point estimate. In general, a smaller value of the *RMSE* suggests that the more accurate the parameter estimate is.

In addition, the mean absolute error (*MAE*) was also used to evaluate the accuracy of item parameter estimates. It is defined as:

$$MAE_{\pi} = \frac{1}{n}\sum_{j=1}^{n}\left|\hat{\pi}_j - \pi_j\right|, \qquad (3.2.5)$$

where π, $\hat{\pi}$, and *n* are defined previously.

Similar to *RMSE*, the smaller the *MAE* is, the more accurate the item parameters are estimated. The *MAE* was considered in addition to *RMSE* because it is intuitively easier to conceptualize, as it measures the absolute difference between the predicted and true values, or the average amount of absolute estimation error in the item.

Each item had a corresponding *bias*, *RMSE*, and *MAE* measures, which was averaged across items to provide summary information. As for the recovery of the person ability parameters, correlations, *r*$(\theta, \hat{\theta})$, was examined between true and estimated values of them. In the educational setting, one is often more interested in the relative values of θ rather than the true values for different examinees. The correlations for both Gibbs sampling and NUTS in various test conditions were averaged across the ten replications and summarized to provide information regarding the accuracy in estimating person abilities.

In addition, in order to determine which factor accounted for most of the variation in the accuracy of estimating of the parameters in the 2PL UIRT model, five separate analyses of variance (ANOVAs) were conducted with the dependent variables being log*RMSE*s and log*MAE*s for estimating the discrimination or difficulty parameters as well as correlations *r*$(\theta, \hat{\theta})$ for estimating the person ability parameters. A log-transformation of the *RMSE*s, log*RMSE*s and the *MAE*s, log*MAE*s were used to increase the likelihood of satisfying the assumption of normality needed in hypothesis testing (Harwell et al., 1996). Effect sizes ($\widehat{\omega}^2$) for the four factors (i.e., sample size (*N*), test length (*K*), prior specifications for a_j (*P*), and MCMC algorithm (*A*)) in each ANOVA were obtained using

$$\widehat{\omega}^2 = \frac{SS_{Effect} - (df_{Effect})(MS_{Error})}{MS_{Error} + SS_{Total}}, \tag{3.2.6}$$

where SS_{Effect} is the sum of squares for a main effect or interaction, df_{Effect} is the degrees of freedom for a main effect or interaction, MS_{Error} is the mean square of the error, and SS_{Total} is the sum of squares for the total model to measure the effect of each main effect and interactions on the respective estimate. Following Cohen’s (1988) guidelines, a large effect captures at least 14% of the variability, a medium effect captures about 6% of the variability, and a small effect is one that captures about 1% of the variance.

3.3 Simulation Study 2

Monte Carlo simulations using Gibbs sampling and NUTS were conducted to answer research question two by examining the recovery of the model parameters using Gibbs sampling and NUTS under various test conditions.

3.3.1 Model

Consider a *K*-item test containing *m* subtests, each consisting of k_v multiple-choice items that measure one ability dimension. With a logit link, the 2PL multi-unidimensional model can be defined as follows:

$$P(Y_{vij} = 1|\theta_{vi}, a_{vj}, b_{vj}) = \frac{\exp[a_{vj}(\theta_{vi}-b_{vj})]}{1+\exp[a_{vj}(\theta_{vi}-b_{vj})]}, \qquad (3.3.1)$$

where Y_{vij} is the response of *i*th individual to *j*th item of *v*th dimension correctly (Y_{vij} = 1) or incorrectly (Y_{vij} = 0), θ_{vi} is the latent ability parameter and is the *v*th component of vector $\boldsymbol{\theta}_{\mathbf{i}}$, a_{vj} is the discrimination parameter of *j*th item of dimension *v*, and b_{vj} is the difficulty parameters of *j*th item of dimension *v*. Note that with each item measuring only one latent ability, the multi-unidimensional IRT model is a special case of the MIRT model.

3.3.2 Simulation Procedure

For the 2PL multi-unidimensional IRT model, tests with two subscales were considered so that the first half items measured one latent trait (θ_1) and the second half measured the other latent trait (θ_2). Three factors were manipulated in the simulation study such that sample size (*N*) was 100, 300, 500, and 1000 examinees, test length (*K*) was 10, 20, and 40 items, and intertrait correlation (ρ_{12}) was 0.2, 0.5, and 0.8 (Sheng & Wikle, 2008). Person parameters $\boldsymbol{\theta}_i$ = $(\theta_{1i}, \theta_{2i})'$ were generated from a bivariate normal distribution such that

$$\boldsymbol{\theta}_i \sim N_2(\boldsymbol{\mu}, \Sigma),$$

where $\boldsymbol{\mu}$ = $(0, 0)'$ and $\sum$ = $\begin{pmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{pmatrix}$. The intertrait ρ_{12} is described previously which was manipulated to be 0.2, 0.5, and 0.8. Item parameters a_{vj} and b_{vj} were generated from uniform distributions such that a_{vj}~$U(0, 2)$, and $b_{vj}{\sim}U(-2, 2)$.

Dichotomous item responses were then generated from the 2PL multi-unidimensional model as defined in Equation (3.3.1), where v = 2.

To implement the 2PL multi-unidimensional model using MCMC, the prior for $\boldsymbol{\theta}_i$ was assumed to follow a multivariate normal distribution

$$\boldsymbol{\theta}_i{\sim}MVN(\boldsymbol{\mu}, \textstyle\sum_H),$$

where $\boldsymbol{\mu}$ = $(0, 0)'$ and $\sum_H$ had an inverse-Wishart prior distribution such that

$$\textstyle\sum_H{\sim}\text{inv-Wishart}(\sum, D),$$

in which D is the degrees of freedom and was set to 2. $\sum$ is a covariance matrix and $\sum$ $= \begin{pmatrix} s_1^2 & \rho_{12}s_1s_2 \\ \rho_{12}s_1s_2 & s_2^2 \end{pmatrix}$,

where s_1, s_2, and ρ_{12} were specified such that s_1~$U(0, 10)$, s_2~$U(0, 10)$, and ρ_{12}~$U(-1, 1)$. Note that the informativeness of the prior for the covariance matrix is decided by D. Also, prior densities for a_{vj} and b_{vj} were set such that $a_{vj}{\sim}N_{(0,\infty)}$ (0, 1) and $b_{vj}{\sim}N(0, 1)$.

Gibbs sampling and NUTS were implemented for each simulated data set via the use of JAGS and Stan, respectively, where the burn-in (or warm-up) stage was set to 3000 iterations followed by 4 chains with 5000 iterations. For both algorithms, the initial values for the discrimination parameters a_{vj} were set to ones, and those for the difficulty parameters b_{vj} and latent ability parameters θ_{vi} were set to zeros. Further, convergence of Markov chains was evaluated using the Gelman-Rubin R statistic (Gelman & Rubin, 1992).

3.3.3 Measures of Estimation Accuracy

For each simulated condition out of the total sample sizes (4) × test lengths (3) × intertrait correlations (3) × algorithms (2) = 72 experimental conditions, ten replications were conducted to avoid erroneous results in estimation due to sampling error. Similar to simulation study 1, the accuracy of item parameter estimates was evaluated using *bias*, *RMSE*, and *MAE*, which are defined in Equations (3.2.3) to (3.2.5). These measures were averaged across items to provide summary information. Also, the recovery of the person ability parameters was examined using correlations $r(\theta_1, \hat{\theta}_1)$ and $r(\theta_2, \hat{\theta}_2)$ between true and estimated values of them. Then, the correlations for both Gibbs sampling and NUTS in various test conditions were averaged across the ten replications and summarized to provide information regarding the accuracy in estimating person abilities.

Similarly, to determine which factor accounted for most of the variation in the accuracy in estimation of the model parameters, ten separate ANOVAs were conducted with dependent variables being log*RMSE*s and log*MAE*s for estimating the discrimination or difficulty parameters in dimension 1 or 2, and correlations $r(\theta_1, \hat{\theta}_1)$ and $r(\theta_2, \hat{\theta}_2)$ for estimating person ability parameters in the two dimensions. For each ANOVA, effect sizes $\hat{\omega}^2$, as defined in Equations (3.2.6), were used to assess the effect of each of the main effects (i.e., sample size (*N*), test length (*K*), intertrait correlation (ρ_{12}), and MCMC algorithm (*A*)) and their interactions.

CHAPTER 4

RESULTS

This chapter summarizes the simulation results comparing Gibbs sampling with NUTS when fitting the two-parameter logistic (2PL) unidimensional IRT (UIRT) and multi-unidimensional IRT models to simulated data. The results are organized such that the first section pertains to findings about the 2PL UIRT model and the second section pertains to findings of the 2PL multi-unidimensional IRT model under different simulated conditions that have been described in Chapter 3.

4.1 Results for the 2PL UIRT Model

As described in Chapter 3, the convergence of Markov chains was examined using the Gelman-Rubin R statistic (Gelman & Rubin, 1992). For the 2PL UIRT model where the burn-in (or warm-up) stage was set to 3000 iterations followed by 4 chains with 5000 iterations, $\hat{R}$ is less than 1.10 for each model parameter under all test conditions using Gibbs sampling or NUTS, suggesting that convergence is potentially achieved using either algorithm. In addition, visual diagnostics of convergence can also be carried out using trace plots and Gelman-Rubin plots as shown in Figures 1 and 2 for one item. With this illustrated item, the trace plots of a_j and b_j using Gibbs sampling (the upper panels) or NUTS (the lower panels) do not demonstrate signs of orphaned chains for item parameters, suggesting that the chains appear to mix well and have converged to the posterior distribution (see Figure 1). Also, the Gelman-Rubin plots show that the Gelman-Rubin R statistic is very close to 1.0 using Gibbs sampling (the upper panels) or NUTS (the lower panels) for item parameters near the end of the sampling period, suggesting that the chains have potentially achieved convergence (see Figure 2).

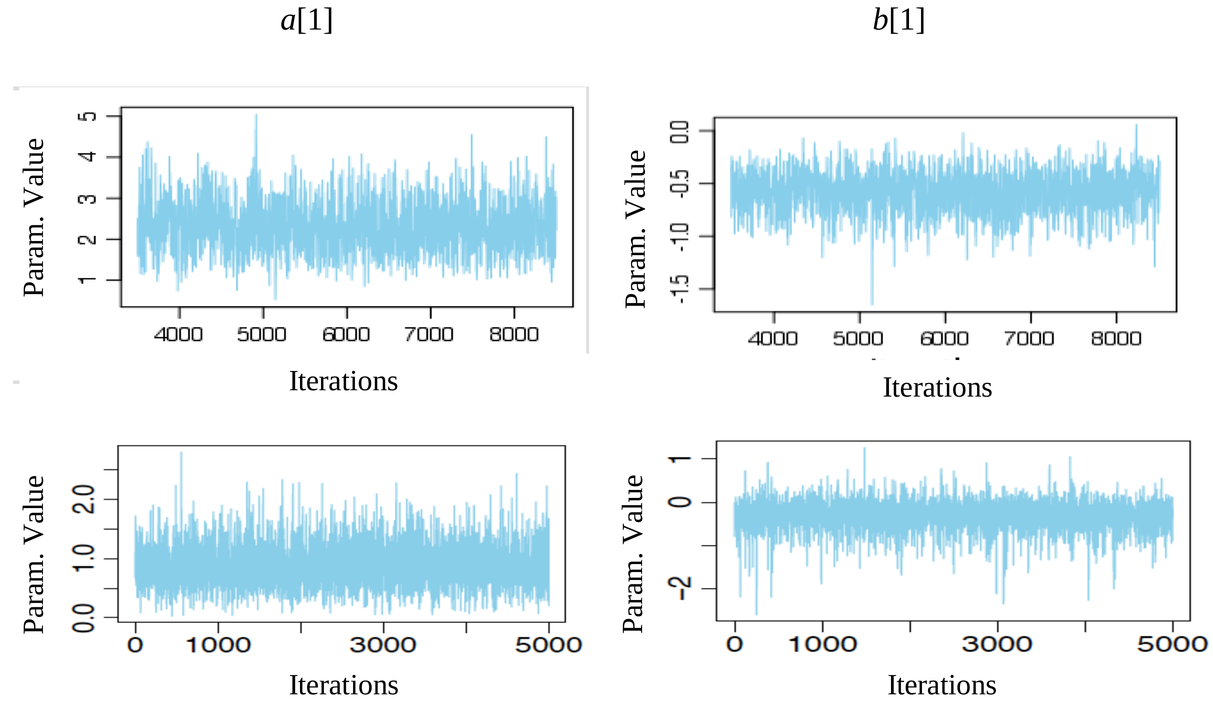

Figure 1. Trace plots of the discrimination parameter and difficulty parameter for one item in the 2PL UIRT model using Gibbs sampling (top) and NUTS (bottom).

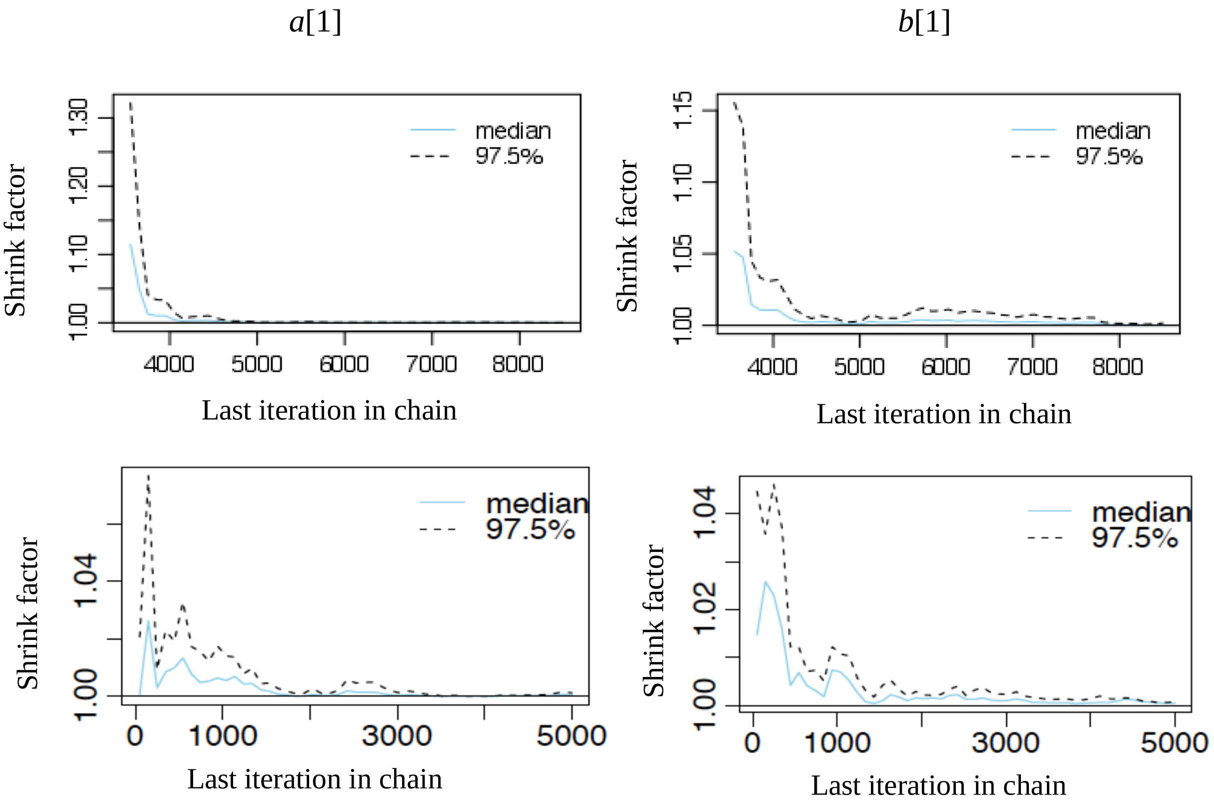

Figure 2. Gelman-Rubin plots of the discrimination parameter and difficulty parameter for one item in the 2PL UIRT model using Gibbs sampling (top) and NUTS (bottom).

4.1.1 Item Parameter Recovery

The average *bias*, root mean square error (*RMSE*), and mean absolute error (*MAE*) values averaged across items for each simulated condition by implementing Gibbs sampling or NUTS to recover the discrimination (a_j) and difficulty (b_j) parameters are summarized in Tables 1 through 4. For visual help, the average *MAE* values to recover the discrimination (a_j) and difficulty (b_j) parameters are summarized in Figures 3 and 4. Since the results of the average *RMSE*s and *MAE*s are similar, only the average *MAE*s are presented in figures.

The results show that Gibbs sampling performs similarly to NUTS under most simulated conditions. Both algorithms recover item parameters with a similar precision as the *RMSE*s and *MAE*s are nearly identical except that they tend to be larger with the maximum value of the *RMSE* equal to 1.639 and *MAE* equal to 0.730 in the condition where the prior distribution for a_j is lognormal using Gibbs sampling with a small sample size (i.e., *N*=100) (see Table 1). With adequate sample sizes and sufficient number of items (e.g., *N*=1000 and *K*=20), discrimination parameter estimates become more stable with the maximum value of the *RMSE* and *MAE* equal to 0.127 and 0.097, respectively (see Table 4). In addition, except for the condition where the lognormal prior is used in Gibbs sampling with *N*=100 and *K*=10 or 20 (see Table 1), *bias* is close to zero for most conditions, suggesting that both algorithms estimate item parameters with little bias.

When sample size increases, the *RMSE*s and *MAE*s for estimating the discrimination tend to decrease using either Gibbs sampling or NUTS (see Figure 3). For example, with the lognormal prior for the discrimination parameter using Gibbs sampling, the *RMSE*s and *MAE*s decrease from 0.862 and 0.373 to 0.127 and 0.094, respectively when *N* increases from 100 to 1000 with *K*=20 (see Tables 1 and 4). Similarly, as sample size increases, the *RMSE*s and *MAE*s for estimating the difficulty parameter tend to decrease using either algorithm except for the condition where the prior specifications for a_j is lognormal with *N*=500 and *K*=20 (see Figure 4). This pattern, however, is not observed with *bias*, which has mixed results. When test length increases, the *RMSE*s and *MAE*s for estimating the discrimination parameter but not the difficulty parameter appear to decrease using either algorithm especially when *N*≥300. For example, with the truncated normal prior for the discrimination parameter using NUTS, the *RMSE*s and *MAE*s decrease from 0.194 and 0.140 to 0.155 and 0.114, respectively when *K* increases from 10 to 40 with *N*=500 (see Table 3). This pattern, however, is not directly observed with *bias*, either.

With both algorithms, the truncated normal prior for the discrimination parameter recovers a_j better than the other two prior specifications when *N*≤300 (see Figure 3). Moreover, with both algorithms, the discrimination parameter tends to recover better than the difficulty parameter for all test conditions except for the condition where the prior for a_j is lognormal using Gibbs sampling with *N*=100, or is exponentiated using either algorithm with *N*=100 and *K*=10. In addition, when comparing the *RMSE*s and *MAE*s under the condition where the prior distribution for a_j is lognormal, there are some cases where the values of the *RMSE*s are slightly larger for Gibbs sampling than those for NUTS, but the values of the *MAE*s are slightly smaller for Gibbs sampling than those for NUTS (see Tables 2 and 3). For example, when the prior distribution for the discrimination parameter is lognormal with *N*=300 and *K*=40, the *RMSE* for estimating a_j using Gibbs sampling is 0.219, which is a little larger than 0.204 using NUTS. The *MAE*, however, for estimating a_j using Gibbs sampling is 0.158, which is a little smaller than 0.164 using NUTS (see Table 2). Although the *RMSE* and *MAE* values are similar, based on the average *RMSE*s, the lognormal prior for the discrimination parameter results in a slightly better estimation for the discrimination parameter when using NUTS with N=300 and K=40. However, based on the *MAE*s, the lognormal prior for the discrimination parameter results in a slightly better estimation when using Gibbs sampling with the same sample size and test length.

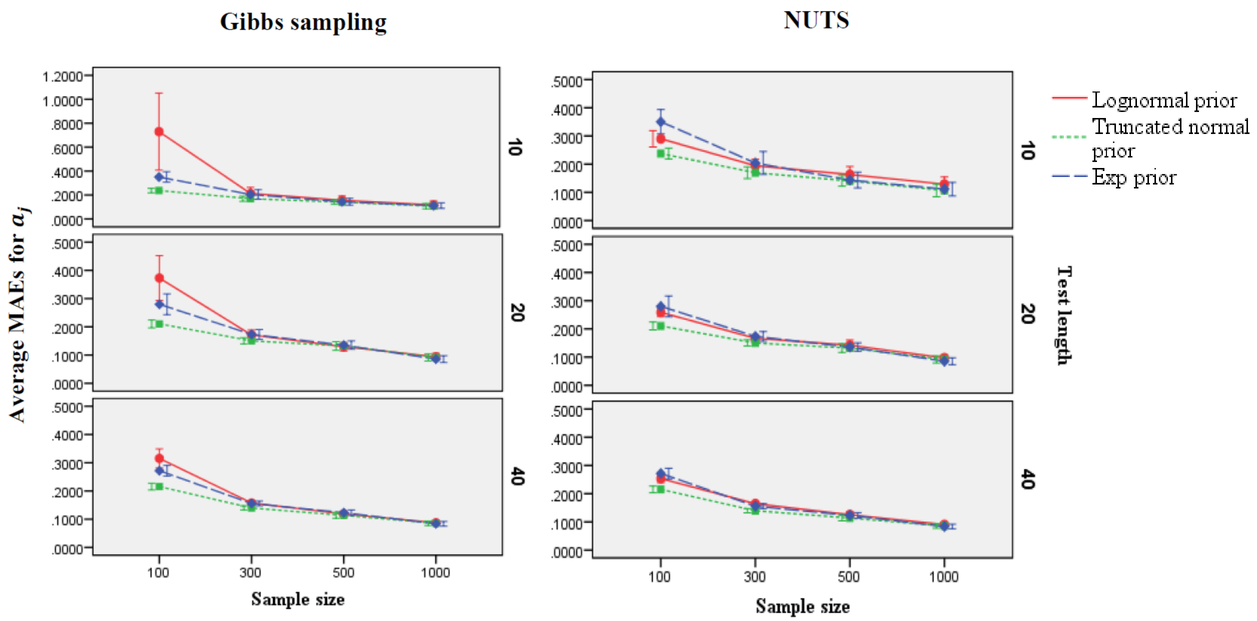

Figure 3. Average MAEs for recovering discrimination (a_j) parameters under various test conditions in the 2PL UIRT model. Note. 25 replications for N=100 and 300, 10 replications for N=500 and 1000.

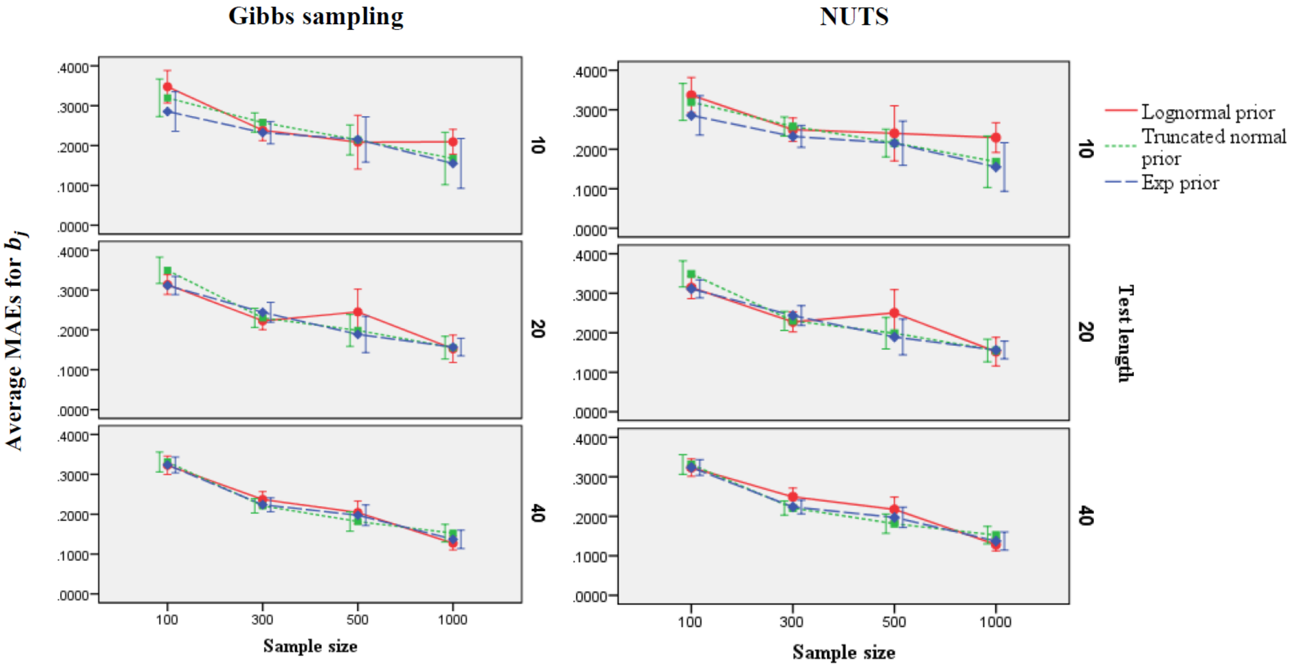

Figure 4. Average MAEs for recovering difficulty (b_j) parameters under various test conditions in the 2PL UIRT model. Note. 25 replications for N=100 and 300, 10 replications for N=500 and 1000.

Table 1. Average Bias, RMSE, and MAE for recovering item parameters in the 2PL UIRT model when N=100.

K	Prior for a_j	Parameters	Gibbs sampling			NUTS		
			Bias[a]	*RMSE*[a]	*MAE*[a]	*Bias*[a]	*RMSE*[a]	*MAE*[a]
10	1	*a*	0.538	1.639	0.730	0.108	0.347	0.290
		b	0.019	0.476	0.348	0.016	0.475	0.337
	2	*a*	-0.001	0.300	0.237	-0.001	0.300	0.237
		b	-0.040	0.466	0.319	-0.039	0.466	0.320
	3	*a*	0.071	0.509	0.351	0.068	0.504	0.350
		b	0.034	0.420	0.285	0.033	0.420	0.286
20	1	*a*	0.209	0.862	0.373	0.105	0.322	0.259
		b	0.018	0.448	0.314	0.017	0.452	0.315
	2	*a*	-0.011	0.273	0.211	-0.011	0.273	0.210
		b	-0.006	0.505	0.349	-0.006	0.505	0.349
	3	*a*	0.080	0.438	0.280	0.080	0.439	0.280
		b	0.009	0.440	0.311	0.008	0.440	0.311
40	1	*a*	0.164	0.554	0.315	0.103	0.313	0.253
		b	0.037	0.462	0.323	0.038	0.469	0.323
	2	*a*	-0.015	0.284	0.215	-0.015	0.284	0.215
		b	-0.009	0.472	0.331	-0.008	0.472	0.331
	3	*a*	0.077	0.393	0.272	0.077	0.392	0.272
		b	-0.030	0.469	0.324	-0.030	0.469	0.324

Note. Prior 1: a_j ~ lognormal(0, 0.5); Prior 2: $a_j \sim N_{(0,\infty)}$ (0, 1); Prior 3: $a_j = \exp(\alpha_j)$, $\alpha_j \sim N(0, 1)$; [a] Based on 25 replications.

Table 2. Average Bias, RMSE, and MAE for recovering item parameters in the 2PL UIRT model when N=300[a].

K	Prior for a_j	Parameters	Gibbs sampling			NUTS		
			Bias[a]	*RMSE*[a]	*MAE*[a]	*Bias*[a]	*RMSE*[a]	*MAE*[a]
10	1	*a*	0.062	0.324	0.213	0.044	0.251	0.194
		b	0.015	0.336	0.238	0.019	0.358	0.249
	2	*a*	0.006	0.227	0.169	0.006	0.228	0.170
		b	-0.005	0.407	0.258	-0.004	0.408	0.258
	3	*a*	0.038	0.292	0.204	0.038	0.292	0.204
		b	0.052	0.340	0.232	0.052	0.341	0.232
20	1	*a*	0.056	0.253	0.171	0.060	0.214	0.167
		b	0.010	0.334	0.223	0.009	0.345	0.228
	2	*a*	0.002	0.199	0.150	0.001	0.199	0.150
		b	0.050	0.368	0.230	0.051	0.367	0.230
	3	*a*	0.032	0.234	0.173	0.032	0.234	0.173
		b	-0.049	0.368	0.244	-0.049	0.367	0.244
40	1	*a*	0.069	0.219	0.158	0.081	0.204	0.164
		b	0.007	0.379	0.237	0.008	0.399	0.249
	2	*a*	-0.005	0.179	0.140	-0.005	0.179	0.140
		b	-0.023	0.341	0.221	-0.023	0.341	0.221
	3	*a*	0.042	0.213	0.156	0.042	0.213	0.156
		b	0.017	0.349	0.224	0.017	0.349	0.224

Note. Prior 1: a_j ~ lognormal(0, 0.5); Prior 2: $a_j \sim N_{(0,\infty)}$ (0, 1); Prior 3: $a_j = \exp(\alpha_j)$, $\alpha_j \sim N(0, 1)$; [a]25 replications.

Table 3. Average Bias, RMSE, and MAE for recovering item parameters in the 2PL UIRT model when *N*=500.

K	Prior for a_j	Parameters	Gibbs sampling			NUTS		
			Bias[a]	*RMSE*[a]	*MAE*[a]	*Bias*[a]	*RMSE*[a]	*MAE*[a]
10	1	*a*	0.019	0.236	0.156	0.023	0.208	0.163
		b	-0.012	0.374	0.208	-0.029	0.410	0.240
	2	*a*	-0.026	0.194	0.141	-0.025	0.194	0.140
		b	0.019	0.307	0.214	0.017	0.310	0.216
	3	*a*	0.018	0.185	0.144	0.016	0.184	0.143
		b	-0.018	0.369	0.215	-0.019	0.368	0.215
20	1	*a*	0.052	0.179	0.130	0.074	0.176	0.142
		b	0.034	0.408	0.245	0.034	0.424	0.250
	2	*a*	0.013	0.184	0.132	0.013	0.184	0.132
		b	-0.006	0.314	0.199	-0.006	0.313	0.199
	3	*a*	-0.007	0.183	0.135	-0.007	0.183	0.136
		b	-0.021	0.326	0.189	-0.022	0.326	0.189
40	1	*a*	0.031	0.158	0.119	0.055	0.159	0.125
		b	-0.034	0.342	0.204	-0.039	0.360	0.217
	2	*a*	0.003	0.154	0.114	0.002	0.155	0.114
		b	-0.028	0.283	0.182	-0.028	0.283	0.181
	3	*a*	0.014	0.161	0.123	0.014	0.161	0.123
		b	0.005	0.327	0.197	0.006	0.327	0.197

Note. Prior 1: a_j ~ lognormal(0, 0.5); Prior 2: a_j ~ $N_{(0,\infty)}$ (0, 1); Prior 3: a_j = exp(α_j), α_j ~ *N*(0, 1); [a] Based on 10 replications.

Table 4. Average Bias, RMSE, and MAE for recovering item parameters in the 2PL UIRT model when *N*=1000.

K	Prior for a_j	Parameters	Gibbs sampling			NUTS		
			Bias[a]	*RMSE*[a]	*MAE*[a]	*Bias*[a]	*RMSE*[a]	*MAE*[a]
10	1	*a*	0.022	0.167	0.117	0.041	0.163	0.129
		b	-0.062	0.312	0.209	-0.081	0.360	0.229
	2	*a*	0.008	0.151	0.107	0.008	0.151	0.107
		b	-0.049	0.302	0.168	-0.050	0.303	0.168
	3	*a*	0.005	0.153	0.111	0.005	0.154	0.111
		b	-0.002	0.287	0.155	-0.003	0.287	0.155
20	1	*a*	0.006	0.127	0.094	0.022	0.127	0.097
		b	-0.0004	0.250	0.153	-0.003	0.253	0.152
	2	*a*	-0.005	0.122	0.091	-0.004	0.121	0.091
		b	0.064	0.280	0.155	0.064	0.280	0.155
	3	*a*	0.026	0.121	0.086	0.026	0.120	0.085
		b	0.006	0.244	0.157	0.006	0.243	0.156
40	1	*a*	0.006	0.117	0.088	0.017	0.118	0.090
		b	-0.025	0.199	0.127	-0.022	0.202	0.128
	2	*a*	-0.006	0.112	0.086	-0.007	0.112	0.086
		b	-0.002	0.274	0.152	-0.002	0.273	0.152
	3	*a*	0.021	0.113	0.084	0.021	0.113	0.084
		b	-0.028	0.226	0.137	-0.029	0.226	0.137

Note. Prior 1: a_j ~ lognormal(0, 0.5); Prior 2: a_j ~ $N_{(0,\infty)}$ (0, 1); Prior 3: a_j = exp(α_j), α_j ~ *N*(0, 1); [a] 10 replications.

4.1.2 Person Ability Parameter Recovery

Correlations between the true and estimated person abilities $r(\theta,\hat{\theta})$ for the 2PL UIRT model are used to evaluate how well the person ability parameters have been recovered under the different simulated conditions using either algorithm and the results are presented in Table 5 and Figure 5. The results show that $r(\theta,\hat{\theta})$ for using both Gibbs sampling and NUTS are nearly the same, indicating that there is not much difference between the two algorithms on estimating the person ability parameter in the 2PL UIRT model. For example, the value of $r(\theta,\hat{\theta})$ is 0.795 using Gibbs sampling and is 0.794 using NUTS when a lognormal prior is assumed for a_j with N=300 and K=10. Furthermore, the choice of priors for the discrimination parameter has a marginal influence on estimating person traits. For example, the values of $r(\theta,\hat{\theta})$ are 0.871, 0.882, and 0.883 using the aforementioned three prior distributions for a_j when NUTS is used with N=500 and K=20. In addition, sample size does not have a considerable effect on the person ability estimates, either. For example, the value of $r(\theta,\hat{\theta})$ remains the same when N increases from 100 to 1000 when a truncated normal prior is assumed for a_j and either algorithm is used with K=40. Test length, however, shows a positive and major effect on estimating the person ability parameter. In other words, the larger K is, the greater $r(\theta,\hat{\theta})$ is, consistent in all simulated conditions, indicating that the person ability parameter is better recovered (see Figure 5). For example, the values of $r(\theta,\hat{\theta})$ change from 0.743 to 0.939 when K increases from 10 to 40 when a lognormal prior is assumed for a_j using Gibbs sampling with N=1000.

In summary, Gibbs sampling and NUTS perform equally well under most of the simulated conditions in estimating the 2PL UIRT model except for the condition where the prior for a_j is lognormal with N=100. In terms of the precision of item parameter estimates, sample size plays a more important role than other test format conditions such as algorithms. Therefore, when other conditions are fixed, more subjects should be added in order to improve the precision of item parameter estimates. On the other hand, when considering the accuracy of person ability parameter estimates, test length plays a crucial role, instead. Therefore, in order to get a better recovery of the person ability parameter, more items should be considered.

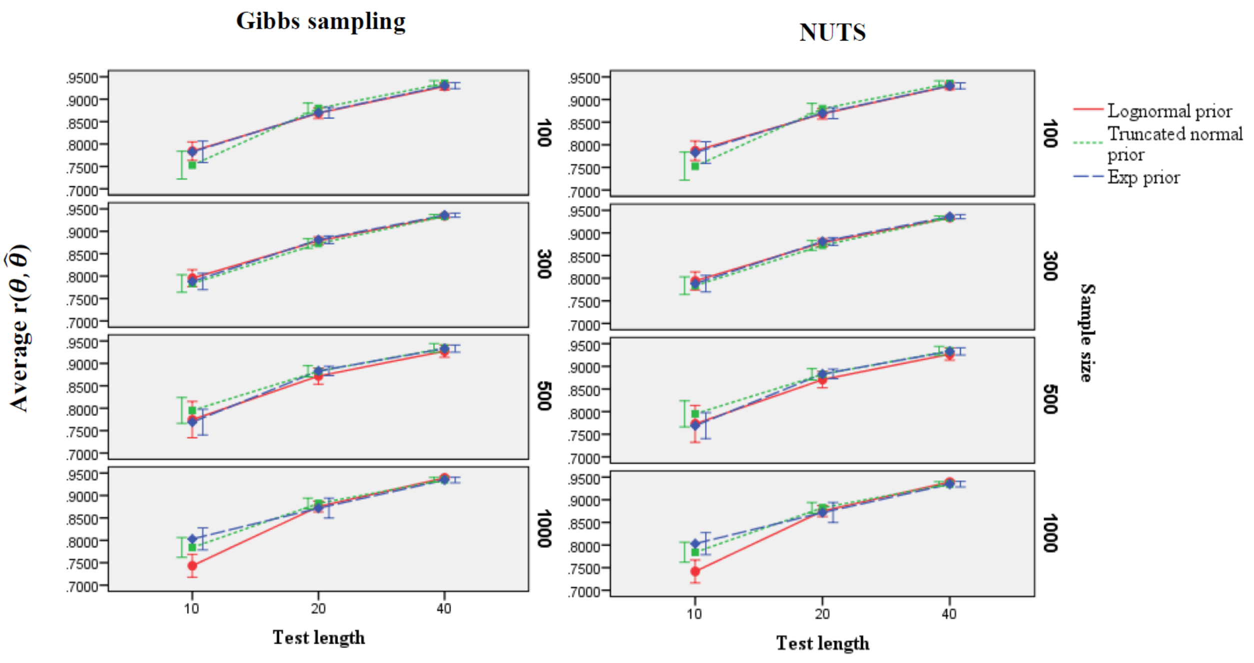

Figure 5. Average correlations between the actual and estimated person abilities $r(\theta,\hat{\theta})$ under various test conditions for the 2PL UIRT model.

Table 5. Correlations between the actual and estimated person abilities $r(\theta,\hat{\theta})$ for the 2PL UIRT model.

N	*K*		Gibbs sampling			NUTS	
		Prior 1	Prior 2	Prior 3	Prior 1	Prior 2	Prior 3
100	10	0.784	0.753	0.782	0.787	0.753	0.783
	20	0.869	0.880	0.870	0.868	0.880	0.870
	40	0.928	0.935	0.930	0.930	0.935	0.930
300	10	0.795	0.783	0.788	0.794	0.783	0.788
	20	0.879	0.873	0.881	0.879	0.873	0.881
	40	0.934	0.933	0.936	0.934	0.933	0.936
500	10	0.774	0.795	0.769	0.773	0.795	0.769
	20	0.872	0.882	0.883	0.871	0.882	0.883
	40	0.927	0.934	0.933	0.927	0.934	0.933
1000	10	0.743	0.784	0.803	0.742	0.784	0.803
	20	0.875	0.882	0.871	0.875	0.882	0.871
	40	0.939	0.935	0.935	0.939	0.935	0.935

Note. Prior 1: a_j ~ lognormal(0, 0.5); Prior 2: $a_j \sim N_{(0,\infty)}$ (0, 1); Prior 3: $a_j = \exp(\alpha_j)$, $\alpha_j \sim N(0, 1)$.

4.1.3 Analysis of Variance (ANOVA) Results for the 2PL UIRT Model

Via the use of ANOVA, effect sizes of the four factors, namely, sample size (*N*), test length (*K*), prior specifications for a_j (*P*) and MCMC algorithm (*A*), in the accuracy of estimating the 2PL UIRT model under various simulated conditions are summarized in Table 6. Also, the interaction plot is presented in Figure 6. Table 6 pertains to the accuracy in estimating the discrimination parameters. It suggests that sample size (*N*) has the largest effect, accounting for 46.2% (49.4%) of the variance in the log*RMSE* (log*MAE*) of the discrimination parameter estimates. Moreover, test length (*K*) has a small effect, accounting for 3.1% (2.9%) of the variance in the log*RMSE* (log*MAE*) of a_j estimates. Prior specifications for a_j (*P*) has a small effect as well, contributing 1.5% (1.5%) of the variance in the log*RMSE* (log*MAE*). In addition, the interaction between sample size (*N*) and prior specifications for a_j (*P*), and interaction among sample size (*N*), prior specifications for a_j (*P*) and algorithm (*A*) have a small effect, accounting for about 1.6% (1.1%) and 1.3% (0.6%), respectively of the variance in the log*RMSE* (log*MAE*) of the discrimination parameter estimates. The interaction plot also suggests that there is an interaction between sample size (*N*) and prior specifications for a_j (*P*) (see Figure 6). This effect indicates that when N≤300, truncated normal prior for a_j should be adopted. However, when N>300, the difference of using these three priors is marginal. The main effect of algorithm (*A*) and other interaction effects are smaller, contributing no more than 1% of the variance in the log*RMSE* (log*MAE*) of a_j estimates.

Regarding to the accuracy in estimating the difficulty parameters, Table 6 indicates that sample size (*N*) again has the largest effect, accounting for about 22.8% (28.5%) in the log*RMSE* (log*MAE*) of the difficulty parameter estimates. However, none of the other main and interaction effects contributes more than 1% of the variance in the log*RMSE* or log*MAE* of the difficulty parameter estimates. For example, test length (*K*) accounts for only 0.6% of the variance in the log*RMSE* of the difficulty parameter estimates.

In terms of the accuracy in estimating the person ability parameters, it shows that with a large effect size, test length (*K*) accounts for the majority of the variance, about 64.0%, in the correlation between θ and $\hat{\theta}$. On the other hand, none of the other main and interaction effects contributes more than 1% of the variance in the correlations, with the maximum value of $\hat{\omega}^2$ equal to 0.007, which is the interaction among sample size (*N*), test length (*K*), and prior specifications for a_j (*P*).

In summary, the ANOVA results reinforce the conclusions drawn from Tables 1 through 5. Increased sample size has a positive and major effect on the recovery of item parameters. Increased test length, on the other hand, positively affects the estimation of the person ability parameters.

Table 6. ANOVA effect sizes ($\hat{\omega}^2$) for log*RMSE* and log*MAE* in estimating the discrimination (*a*), difficulty (*b*) parameters, and $r(\theta,\hat{\theta})$ in the 2PL UIRT model.

Variable	log*RMSEa*	log*MAEa*	log*RMSEb*	log*MAEb*	$r(\theta,\hat{\theta})$
P	0.015	0.015	0.002	0.003	0.000
K	0.031	0.029	0.006	0.007	0.640
N	0.462	0.494	0.228	0.285	0.001
A	0.003	0.000	0.000	0.000	0.000
P×*K*	0.000	0.001	0.001	0.001	0.000
P×*N*	0.016	0.011	0.001	0.000	0.002
P×*A*	0.005	0.001	0.000	0.000	0.000
K×*N*	0.001	0.001	0.003	0.004	0.000
K×*A*	0.000	0.000	0.000	0.000	0.000
N×*A*	0.006	0.003	0.000	0.000	0.000
P×*K*×*N*	0.002	0.000	0.010	0.008	0.007
P×*K*×*A*	0.001	0.000	0.000	0.000	0.000
P×*N*×*A*	0.013	0.006	0.000	0.000	0.000
K×*N*×*A*	0.001	0.000	0.000	0.000	0.000
P×*K*×*N*×*A*	0.001	0.000	0.000	0.000	0.000

Note. *P*: Prior specifications for a_j; *K*: Test length; *N*: Sample size; *A*: Algorithm.

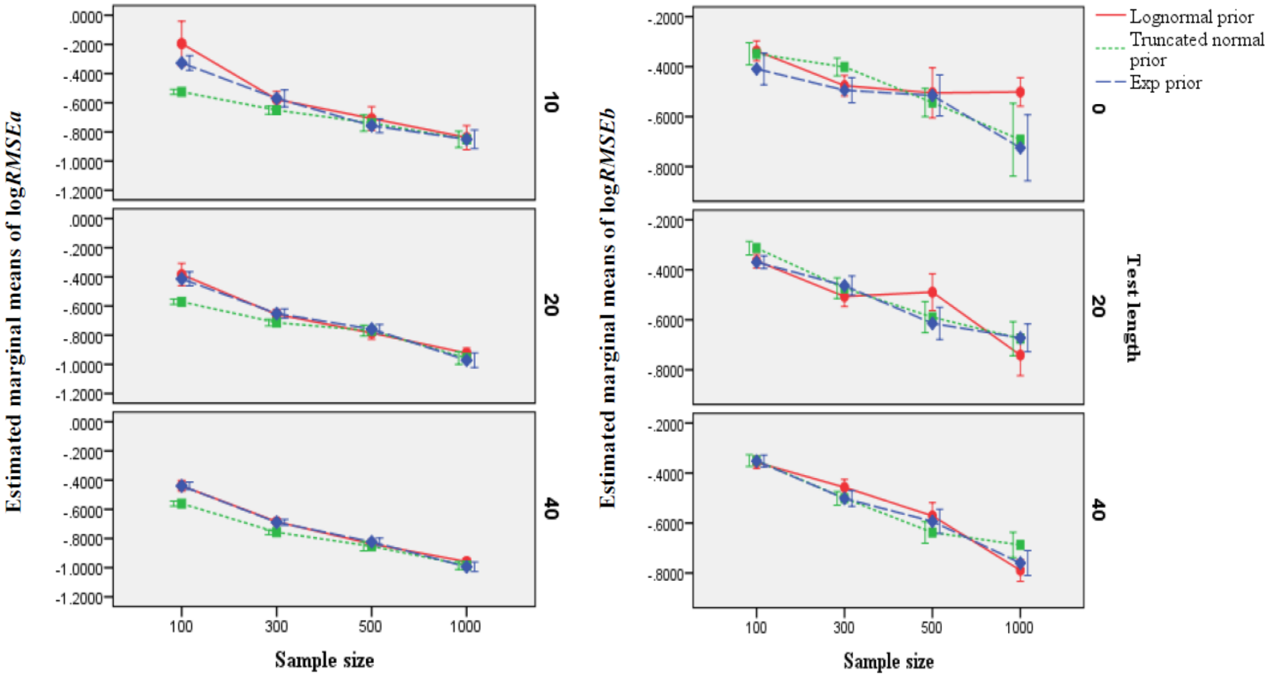

Figure 6. The interactions of sample size and prior specifications for a_j under different test lengths in logRMSEa (left) and logRMSEb (right).

4.2 Results for the 2PL Multi-unidimensional IRT Model

For the 2PL multi-unidimensional IRT model where the burn-in (or warm-up) stage was set to 3000 iterations followed by 4 chains with 5000 iterations, $\hat{R}$ is also less than 1.10 for each model parameter under all simulated conditions using either Gibbs sampling or NUTS, suggesting that convergence is potentially achieved using either algorithm.

4.2.1 Item Parameter Recovery

The average *bias*, *RMSE*, and *MAE* values averaged across items for recovering the discrimination (a_1, a_2), difficulty (b_1, b_2), and intertrait correlation (ρ_{12}) parameters in the 2PL multi-unidimensional model via the use of Gibbs sampling and NUTS are summarized in Tables 7 through 10. For visual help, the average *MAE* values to recover the discrimination (a_1, a_2) and difficulty (b_1, b_2) parameters are summarized in Figures 7 through 10. For all item parameters,

a_1 and b_1 are used to denote the discrimination and difficulty parameters for the first half items, which are assumed to measure θ_1, and a_2 and b_2 are used to denote the discrimination and difficulty parameters for the second half, which are assumed to measure θ_2.

A close examination of the tables suggests that Gibbs sampling and NUTS do not differ much in their average *bias*, *RMSE*, and *MAE* values in estimating individual item parameters. In addition, both algorithms recover the intertrait correlation (ρ_{12}) parameter with a similar precision as the *RMSE*s and *MAE*s are nearly identical. As the sample size increases, the *RMSE*s and *MAE*s for the discrimination parameters (a_1, a_2) tend to decrease using either Gibbs sampling or NUTS (see Figures 7 and 8). For example, when *N* increases from 100 to 1000, the *RMSE*s and *MAE*s decrease from 0.277 and 0.210 to 0.171 and 0.128, respectively for a_1 using Gibbs sampling with ρ_{12}= 0.5 and *K*=20 (see Tables 7 and 10). This pattern, however, is not observed with *RMSE*s (*MAE*s) for the difficulty parameters (b_1, b_2) (see Figures 9 and 10) and *bias*. As the test length increases, the *RMSE*s and *MAE*s for estimating the discrimination parameters (a_1, a_2) and difficulty parameters (b_1, b_2) do not show a consistent pattern using either algorithm. Similarly, there is no pattern observed with *bias*. In other words, sample size plays a more crucial role than test length in improving the precision of the discrimination parameter estimates. This is different from what was observed with the simpler 2PL UIRT model where as sample size increases, both discrimination and difficulty parameters are estimated more accurately.

In terms of recovering the intertrait correlation (ρ_{12}) parameter, the results indicate that sample size or test length does not show a consistent pattern in the accuracy or bias in estimating it. However, when comparing among the various levels of ρ_{12} considered (i.e., 0.2, 0.5, 0.8), the *RMSE*s and *MAE*s for estimating ρ_{12} tend to increase as the actual correlation increases using either algorithm except for the conditions where *N*=100 and *K*≤20. For example, when ρ_{12} increases from 0.2 to 0.8, the *RMSE*s and *MAE*s for estimating the intertrait correlation increase from 0.084 and 0.078 to 0.220 and 0.210, respectively when Gibbs sampling is used with *N*=1000 and *K*=10 (see Table 10). Moreover, the negative average bias values indicate that ρ_{12} is consistently underestimated under all simulated conditions.

In addition, with both algorithms, the discrimination parameter tends to recover better than the difficulty parameter under all simulated conditions except for the condition where ρ_{12}= 0.2 is used together with a small sample size and a short test length (i.e., *N*=100 and *K*=10) (see Table 7).

When comparing the average *RMSE*, *MAE* or bias values for estimating the discrimination parameters (a_1, a_2) and difficulty parameters (b_1, b_2) from different dimensions (i.e., a_1 vs. a_2 and b_1 vs. b_2) under various test conditions, the results indicate that there is an inconsistent pattern.

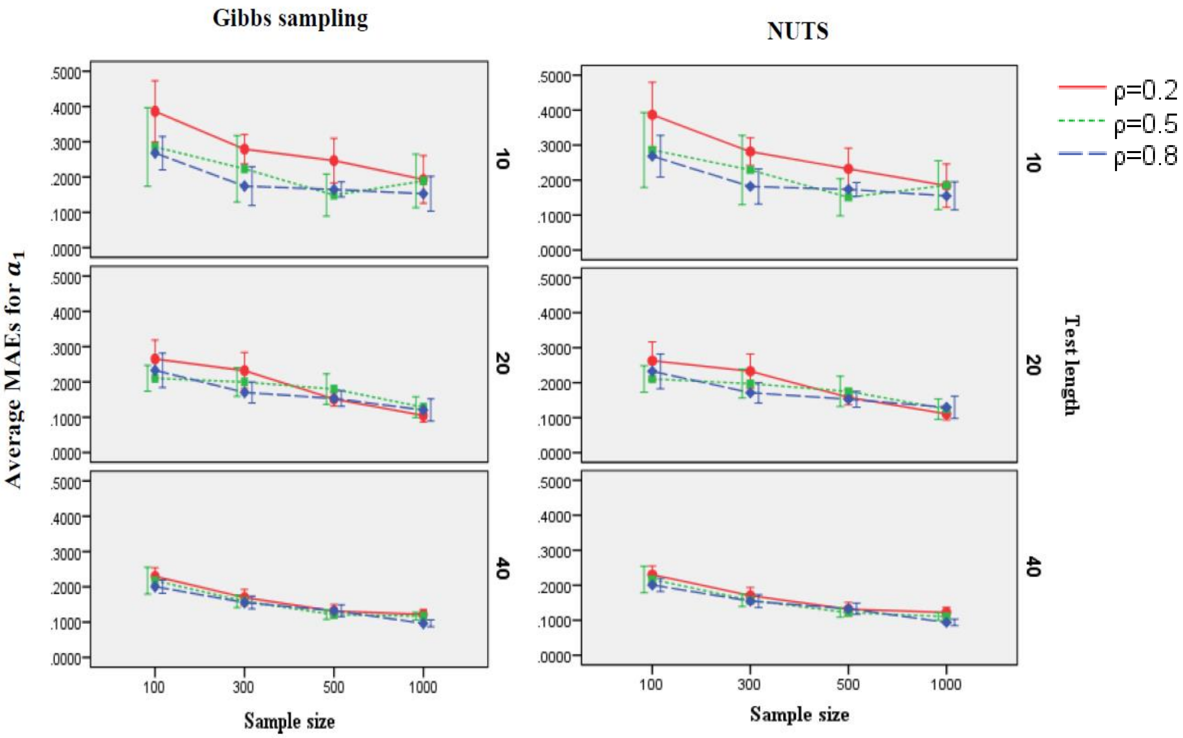

Figure 7. Average MAEs for recovering discrimination (a_1) parameters under various test conditions in the 2PL multi-unidimensional IRT model.

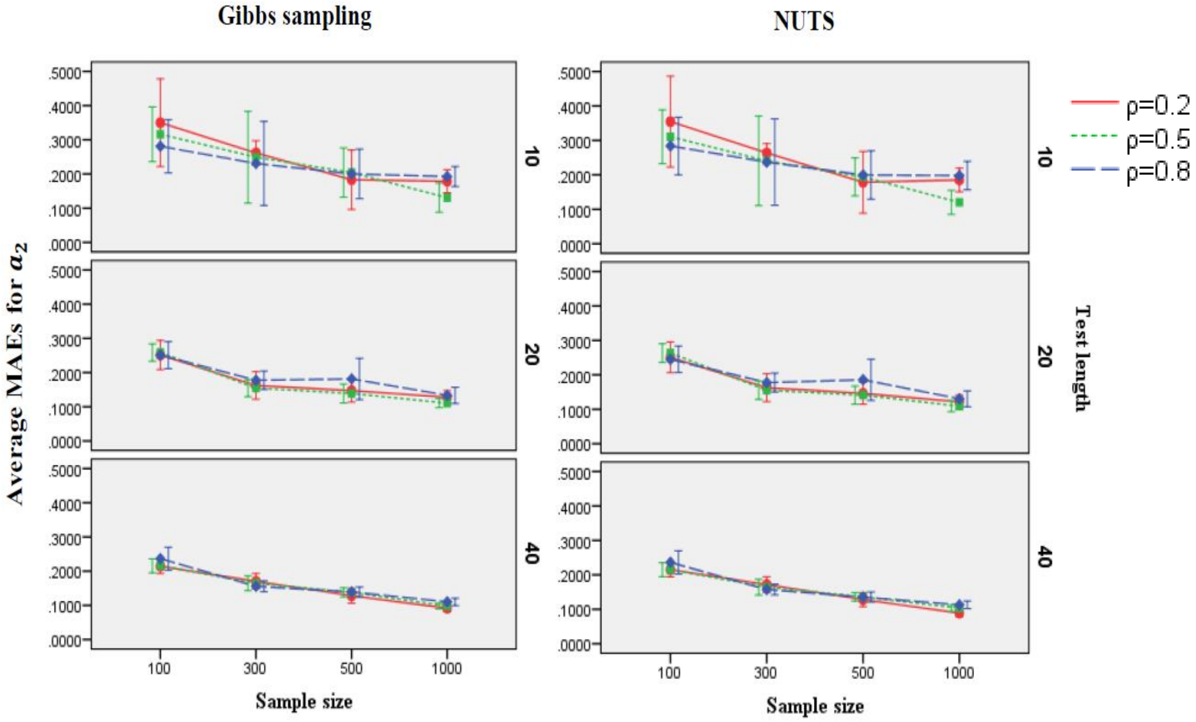

Figure 8. Average MAEs for recovering discrimination (a_2) parameters under various test conditions in the 2PL multi-unidimensional IRT model.

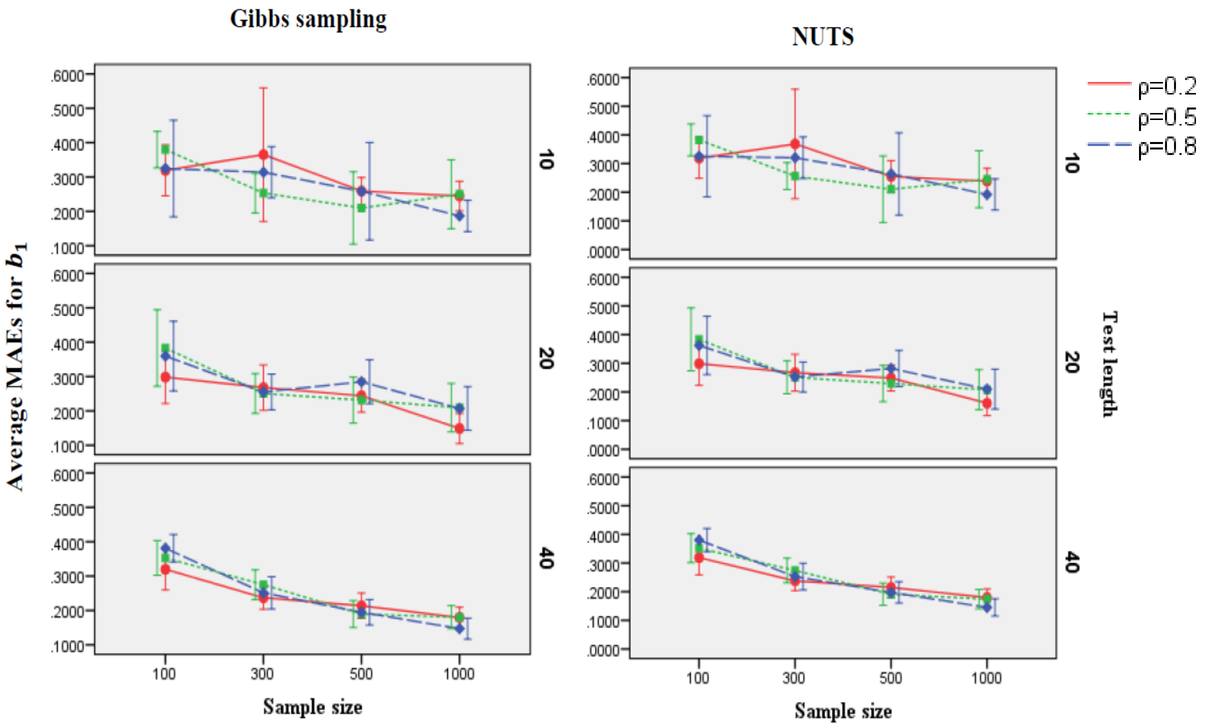

Figure 9. Average MAEs for recovering difficulty (b_1) parameters under various test conditions in the 2PL multi-unidimensional IRT model.

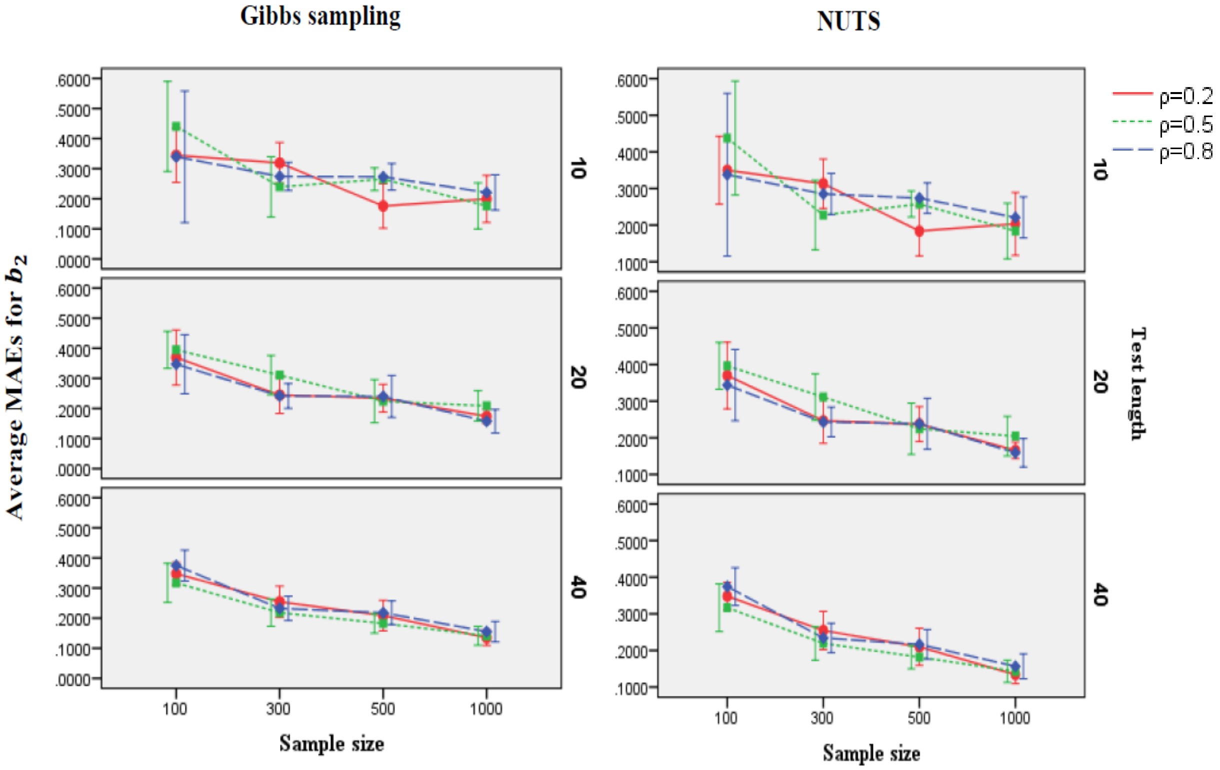

Figure 10. Average MAEs for recovering difficulty (b_2) parameters under various test conditions in the 2PL multi-unidimensional IRT model.

Table 7. Average Bias, RMSE, and MAE for recovering item parameters in the 2PL multi-unidimensional IRT model when *N*=100.

K	ρ_{12}	Parameters	Gibbs sampling			NUTS		
			Bias	*RMSE*	*MAE*	*Bias*	*RMSE*	*MAE*
10	0.2	a_1	-0.055	0.460	0.386	-0.055	0.461	0.386
		b_1	-0.057	0.441	0.319	-0.059	0.444	0.319
		a_2	-0.185	0.445	0.350	-0.189	0.447	0.355
		b_2	0.136	0.435	0.345	0.143	0.441	0.350
		ρ_{12}	-0.071	0.201	0.163	-0.058	0.177	0.138
	0.5	a_1	-0.076	0.355	0.285	-0.071	0.359	0.286
		b_1	0.076	0.490	0.380	0.074	0.492	0.383
		a_2	-0.062	0.379	0.316	-0.061	0.374	0.311
		b_2	-0.094	0.587	0.440	-0.097	0.585	0.438
		ρ_{12}	-0.103	0.149	0.125	-0.099	0.142	0.119
	0.8	a_1	-0.081	0.355	0.268	-0.070	0.355	0.268
		b_1	0.029	0.432	0.324	0.023	0.434	0.325
		a_2	-0.166	0.371	0.281	-0.167	0.371	0.284
		b_2	-0.066	0.522	0.339	-0.070	0.521	0.338
		ρ_{12}	-0.279	0.302	0.279	-0.285	0.306	0.285
20	0.2	a_1	-0.038	0.344	0.265	-0.042	0.342	0.263
		b_1	-0.001	0.430	0.298	-0.001	0.429	0.299
		a_2	-0.066	0.316	0.251	-0.068	0.317	0.251
		b_2	0.018	0.534	0.369	0.019	0.533	0.370
		ρ_{12}	-0.119	0.434	0.119	-0.116	0.432	0.116
	0.5	a_1	0.008	0.277	0.210	0.010	0.276	0.211
		b_1	0.005	0.544	0.383	0.006	0.543	0.384
		a_2	0.023	0.312	0.259	0.027	0.316	0.263
		b_2	-0.007	0.553	0.395	-0.009	0.554	0.396
		ρ_{12}	-0.206	0.782	0.217	-0.200	0.761	0.210
	0.8	a_1	0.009	0.318	0.232	0.003	0.320	0.232
		b_1	-0.080	0.522	0.359	-0.081	0.525	0.362
		a_2	-0.021	0.307	0.251	-0.019	0.299	0.245
		b_2	-0.069	0.495	0.347	-0.067	0.494	0.344
		ρ_{12}	-0.176	0.640	0.176	-0.170	0.627	0.170
40	0.2	a_1	-0.052	0.292	0.229	-0.053	0.293	0.230
		b_1	-0.030	0.446	0.320	-0.029	0.445	0.319
		a_2	-0.030	0.271	0.214	-0.033	0.272	0.215
		b_2	-0.012	0.480	0.348	-0.012	0.481	0.348
		ρ_{12}	-0.125	0.450	0.142	-0.127	0.454	0.143
	0.5	a_1	-0.007	0.287	0.217	-0.005	0.286	0.217
		b_1	0.013	0.500	0.353	0.016	0.499	0.352
		a_2	-0.011	0.272	0.216	-0.013	0.272	0.215
		b_2	0.003	0.457	0.317	0.002	0.457	0.317
		ρ_{12}	-0.170	0.549	0.170	-0.168	0.548	0.168
	0.8	a_1	0.039	0.254	0.200	0.036	0.254	0.201
		b_1	-0.049	0.540	0.381	-0.050	0.538	0.380
		a_2	-0.064	0.298	0.237	-0.063	0.297	0.236
		b_2	0.004	0.526	0.374	0.003	0.526	0.374
		ρ_{12}	-0.208	0.590	0.208	-0.212	0.604	0.212

Table 8. Average Bias, RMSE, and MAE for recovering item parameters in the 2PL multi-unidimensional IRT model when *N*=300.

K	ρ_{12}	Parameters	Gibbs sampling			NUTS		
			Bias	*RMSE*	*MAE*	*Bias*	*RMSE*	*MAE*
10	0.2	a_1	-0.082	0.339	0.279	-0.089	0.344	0.281
		b_1	-0.087	0.518	0.365	-0.088	0.517	0.368
		a_2	-0.055	0.334	0.262	-0.057	0.336	0.264
		b_2	0.040	0.417	0.319	0.035	0.413	0.313
		ρ_{12}	-0.034	0.085	0.072	-0.032	0.082	0.070
	0.5	a_1	-0.058	0.287	0.223	-0.042	0.288	0.229
		b_1	0.102	0.349	0.253	0.106	0.346	0.256
		a_2	-0.079	0.342	0.249	-0.076	0.331	0.241
		b_2	-0.006	0.333	0.240	0.001	0.318	0.228
		ρ_{12}	-0.146	0.170	0.150	-0.139	0.156	0.139
	0.8	a_1	-0.032	0.224	0.174	-0.045	0.238	0.182
		b_1	-0.051	0.445	0.314	-0.054	0.457	0.320
		a_2	-0.034	0.314	0.231	-0.013	0.318	0.237
		b_2	-0.053	0.380	0.274	-0.040	0.396	0.285
		ρ_{12}	-0.178	0.200	0.178	-0.192	0.218	0.192
20	0.2	a_1	-0.026	0.287	0.232	-0.026	0.287	0.233
		b_1	0.003	0.370	0.268	0.003	0.370	0.267
		a_2	-0.025	0.213	0.162	-0.031	0.215	0.163
		b_2	0.001	0.353	0.244	0.001	0.355	0.246
		ρ_{12}	-0.095	0.325	0.099	-0.093	0.321	0.097
	0.5	a_1	-0.074	0.256	0.200	-0.073	0.254	0.197
		b_1	0.024	0.335	0.250	0.024	0.334	0.251
		a_2	0.001	0.209	0.154	-0.001	0.211	0.155
		b_2	-0.043	0.447	0.311	-0.042	0.448	0.311
		ρ_{12}	-0.160	0.451	0.160	-0.154	0.449	0.154
	0.8	a_1	-0.015	0.227	0.170	-0.006	0.228	0.171
		b_1	0.027	0.381	0.255	0.028	0.381	0.252
		a_2	0.023	0.229	0.178	0.016	0.228	0.178
		b_2	-0.010	0.315	0.242	-0.009	0.314	0.243
		ρ_{12}	-0.250	0.665	0.250	-0.246	0.649	0.246
40	0.2	a_1	-0.028	0.231	0.170	-0.031	0.232	0.170
		b_1	-0.025	0.338	0.237	-0.026	0.339	0.238
		a_2	0.011	0.228	0.170	0.009	0.229	0.171
		b_2	0.011	0.398	0.254	0.012	0.398	0.254
		ρ_{12}	-0.073	0.215	0.073	-0.075	0.220	0.075
	0.5	a_1	0.008	0.209	0.159	0.008	0.206	0.157
		b_1	0.023	0.423	0.275	0.022	0.421	0.274
		a_2	-0.026	0.217	0.165	-0.028	0.217	0.164
		b_2	0.046	0.361	0.218	0.045	0.361	0.219
		ρ_{12}	-0.186	0.501	0.186	-0.186	0.496	0.186
	0.8	a_1	0.016	0.201	0.156	0.015	0.200	0.155
		b_1	0.018	0.365	0.251	0.019	0.366	0.253
		a_2	0.006	0.209	0.156	0.005	0.211	0.157
		b_2	-0.001	0.335	0.232	-0.001	0.337	0.234
		ρ_{12}	-0.227	0.602	0.227	-0.225	0.593	0.225

Table 9. Average Bias, RMSE, and MAE for recovering item parameters in the 2PL multi-unidimensional IRT model when *N*=500.

K	ρ_{12}	Parameters	Gibbs sampling			NUTS		
			Bias	*RMSE*	*MAE*	*Bias*	*RMSE*	*MAE*
10	0.2	a_1	-0.061	0.310	0.247	-0.050	0.299	0.232
		b_1	-0.034	0.336	0.259	-0.039	0.330	0.255
		a_2	-0.055	0.241	0.183	-0.077	0.239	0.178
		b_2	0.026	0.244	0.176	0.032	0.251	0.184
		ρ_{12}	-0.110	0.121	0.110	-0.110	0.121	0.110
	0.5	a_1	0.073	0.194	0.149	0.083	0.195	0.151
		b_1	-0.052	0.340	0.210	-0.056	0.343	0.210
		a_2	-0.117	0.247	0.204	-0.111	0.232	0.194
		b_2	-0.001	0.349	0.265	-0.005	0.346	0.258
		ρ_{12}	-0.176	0.190	0.176	-0.177	0.192	0.177
	0.8	a_1	-0.107	0.230	0.165	-0.093	0.234	0.173
		b_1	-0.012	0.475	0.258	-0.014	0.479	0.264
		a_2	-0.104	0.275	0.200	-0.105	0.275	0.200
		b_2	0.046	0.384	0.273	0.046	0.380	0.274
		ρ_{12}	-0.226	0.233	0.226	-0.216	0.225	0.216
20	0.2	a_1	-0.011	0.203	0.151	-0.007	0.212	0.158
		b_1	0.001	0.325	0.244	0.005	0.328	0.249
		a_2	0.015	0.189	0.147	0.008	0.192	0.146
		b_2	-0.049	0.335	0.234	-0.052	0.336	0.237
		ρ_{12}	-0.070	0.208	0.070	-0.071	0.215	0.071
	0.5	a_1	-0.026	0.237	0.180	-0.023	0.234	0.175
		b_1	0.032	0.392	0.232	0.030	0.389	0.229
		a_2	0.030	0.177	0.139	0.027	0.178	0.141
		b_2	-0.043	0.350	0.224	-0.043	0.350	0.225
		ρ_{12}	-0.189	0.538	0.189	-0.188	0.537	0.188
	0.8	a_1	-0.029	0.200	0.153	-0.029	0.199	0.153
		b_1	-0.070	0.444	0.285	-0.071	0.441	0.282
		a_2	0.040	0.249	0.181	0.051	0.255	0.186
		b_2	0.029	0.354	0.240	0.029	0.355	0.238
		ρ_{12}	-0.225	0.675	0.225	-0.221	0.663	0.221
40	0.2	a_1	-0.0004	0.177	0.131	-0.0006	0.177	0.132
		b_1	0.031	0.345	0.214	0.031	0.345	0.215
		a_2	0.018	0.167	0.128	0.023	0.166	0.128
		b_2	0.011	0.318	0.208	0.010	0.320	0.210
		ρ_{12}	-0.066	0.215	0.066	-0.067	0.214	0.067
	0.5	a_1	-0.010	0.161	0.121	-0.019	0.161	0.122
		b_1	-0.034	0.293	0.190	-0.035	0.292	0.191
		a_2	-0.027	0.169	0.138	-0.027	0.167	0.136
		b_2	-0.027	0.288	0.183	-0.027	0.287	0.182
		ρ_{12}	-0.189	0.531	0.189	-0.185	0.526	0.185
	0.8	a_1	-0.009	0.174	0.132	-0.013	0.176	0.133
		b_1	0.003	0.304	0.195	0.004	0.306	0.197
		a_2	-0.030	0.181	0.139	-0.027	0.176	0.135
		b_2	-0.022	0.333	0.218	-0.022	0.332	0.216
		ρ_{12}	-0.247	0.643	0.247	-0.243	0.638	0.243

Table 10. Average Bias, RMSE, and MAE for recovering item parameters in the 2PL multi-unidimensional IRT model when *N*=1000.

K	ρ_{12}	Parameters	Gibbs sampling			NUTS		
			Bias	*RMSE*	*MAE*	*Bias*	*RMSE*	*MAE*
10	0.2	a_1	0.032	0.259	0.193	0.054	0.241	0.184
		b_1	-0.048	0.316	0.244	-0.040	0.317	0.238
		a_2	-0.047	0.224	0.179	-0.054	0.224	0.185
		b_2	0.008	0.266	0.200	-0.0007	0.273	0.203
		ρ_{12}	-0.052	0.084	0.078	-0.045	0.080	0.071
	0.5	a_1	-0.041	0.258	0.189	-0.051	0.259	0.185
		b_1	-0.028	0.359	0.250	-0.020	0.352	0.245
		a_2	0.024	0.164	0.131	-0.004	0.150	0.120
		b_2	-0.039	0.247	0.176	-0.053	0.252	0.184
		ρ_{12}	-0.171	0.180	0.171	-0.169	0.178	0.169
	0.8	a_1	-0.044	0.206	0.153	-0.044	0.214	0.155
		b_1	-0.005	0.234	0.187	-0.012	0.243	0.192
		a_2	-0.103	0.261	0.193	-0.090	0.264	0.198
		b_2	0.012	0.322	0.221	0.011	0.316	0.221
		ρ_{12}	-0.210	0.220	0.210	-0.211	0.219	0.211
20	0.2	a_1	-0.015	0.143	0.105	-0.016	0.148	0.111
		b_1	0.028	0.256	0.149	0.027	0.265	0.161
		a_2	0.034	0.159	0.128	0.025	0.154	0.121
		b_2	-0.011	0.247	0.174	-0.009	0.238	0.166
		ρ_{12}	-0.069	0.217	0.069	-0.069	0.217	0.069
	0.5	a_1	-0.020	0.171	0.128	-0.018	0.166	0.125
		b_1	0.005	0.343	0.209	0.005	0.342	0.208
		a_2	-0.020	0.147	0.111	-0.013	0.145	0.109
		b_2	-0.082	0.321	0.209	-0.083	0.317	0.205
		ρ_{12}	-0.183	0.492	0.183	-0.184	0.497	0.184
	0.8	a_1	0.067	0.178	0.120	0.080	0.185	0.130
		b_1	0.032	0.391	0.207	0.035	0.393	0.210
		a_2	-0.072	0.174	0.134	-0.075	0.171	0.131
		b_2	0.016	0.228	0.157	0.021	0.227	0.159
		ρ_{12}	-0.207	0.554	0.207	-0.204	0.550	0.204
40	0.2	a_1	0.014	0.171	0.122	0.015	0.172	0.122
		b_1	0.002	0.292	0.179	0.003	0.293	0.180
		a_2	-0.008	0.120	0.092	-0.013	0.116	0.089
		b_2	0.009	0.226	0.135	0.010	0.224	0.134
		ρ_{12}	-0.099	0.270	0.099	-0.098	0.269	0.098
	0.5	a_1	-0.012	0.162	0.117	-0.015	0.153	0.111
		b_1	0.015	0.292	0.180	0.011	0.288	0.174
		a_2	0.007	0.134	0.098	0.011	0.139	0.104
		b_2	-0.023	0.259	0.141	-0.023	0.260	0.143
		ρ_{12}	-0.178	0.491	0.178	-0.175	0.487	0.175
	0.8	a_1	-0.030	0.127	0.096	-0.027	0.124	0.094
		b_1	0.010	0.236	0.147	0.011	0.234	0.145
		a_2	-0.034	0.141	0.110	-0.046	0.145	0.113
		b_2	-0.027	0.251	0.155	-0.025	0.252	0.157
		ρ_{12}	-0.215	0.593	0.215	-0.214	0.587	0.214

4.2.2 Person Ability Parameter Recovery

In order to understand the performance of the two algorithms in estimating person abilities in the 2PL multi-unidimensional model, correlations between the true and estimated person abilities from dimension 1, $r(\theta_1,\hat{\theta}_1)$ and dimension 2, $r(\theta_2,\hat{\theta}_2)$ are obtained and presented in Table 11, and Figures 11 and 12.

Consistent to what is observed in the item parameter recovery for this model, $r(\theta_1,\hat{\theta}_1)$ and $r(\theta_2,\hat{\theta}_2)$ do not show much difference between Gibbs sampling and NUTS in estimating the person ability parameter as the values of correlating θ_1 with $\hat{\theta}_1$ or θ_2 with $\hat{\theta}_2$ are almost identical using the two algorithms. Moreover, similar to the 2PL UIRT model, sample size has not much influence on estimating person trait levels, but test length has a positive and major effect on estimating both θ_1 and θ_2. Specifically, with an increase of *K*, $r(\theta_1,\hat{\theta}_1)$ and $r(\theta_2,\hat{\theta}_2)$ consistently increase regardless of *N*, ρ_{12}, or algorithm, suggesting that the person ability parameters are better recovered with more items (see Figures 11 and 12). For example, when *K* increases from 10 to 40, the values of $r(\theta_1,\hat{\theta}_1)$ and $r(\theta_2,\hat{\theta}_2)$ change from 0.673 and 0.690 to 0.879 and 0.891, respectively using NUTS with ρ_{12}=0.5 and *N*=300. In addition to test length, increase of the intertrait correlation also plays a role in estimating the person ability parameters: with a higher intertrait correlation between the two dimensions, the estimated person ability parameters appear to be closer to their true values than conditions with a lower intertrait correlation. For example, when ρ_{12} increases from 0.2 to 0.8, the values of $r(\theta_1,\hat{\theta}_1)$ and $r(\theta_2,\hat{\theta}_2)$ increase from 0.634 and 0.648 to 0.740 and 0.747, respectively using Gibbs sampling with *N*=300 and *K*=10. Also, the effect of test length on the person ability parameter estimation is not the same across the three levels of intertrait correlations. Specifically, a low intertrait correlation seems to benefit more from an increased test length than a high intertrait correlation. For example, when test length increases from 10 to 40, the values of $r(\theta_1,\hat{\theta}_1)$ for ρ_{12}=0.2 and ρ_{12}=0.8 increase from 0.631 and 0.734 to 0.872 and 0.886, respectively using Gibbs sampling with *N*=100. On the other hand, the values of $r(\theta_1,\hat{\theta}_1)$ remain similar across the three levels of intertrait correlations for *K*=40, but increase from 0.631 to 0.734 as ρ_{12} increases from 0.2 to 0.8 for *K*=10 using Gibbs sampling with *N*=100. In addition, when comparing $r(\theta_1,\hat{\theta}_1)$ with $r(\theta_2,\hat{\theta}_2)$, there is not much difference in their values and no specific pattern between them under different simulated conditions especially when *K* is larger (i.e., *K*=40). For example, the values of $r(\theta_1,\hat{\theta}_1)$ and $r(\theta_2,\hat{\theta}_2)$ are the same using Gibbs sampling with *N*=1000, *K*=40, and ρ_{12}=0.5.

In summary, Gibbs sampling and NUTS perform similarly under all of the simulated conditions for estimating the 2PL multi-unidimensional IRT model. The recovery of item parameters shows the pattern that if sample size increases, the precision of the discrimination parameters improves accordingly. Likewise, the recovery of person ability parameters has the pattern that if test length increases, the precision of the person ability parameter estimate becomes better. In addition, test conditions with highly correlated dimensions can also achieve improved precision in the recovery of the person ability parameters.

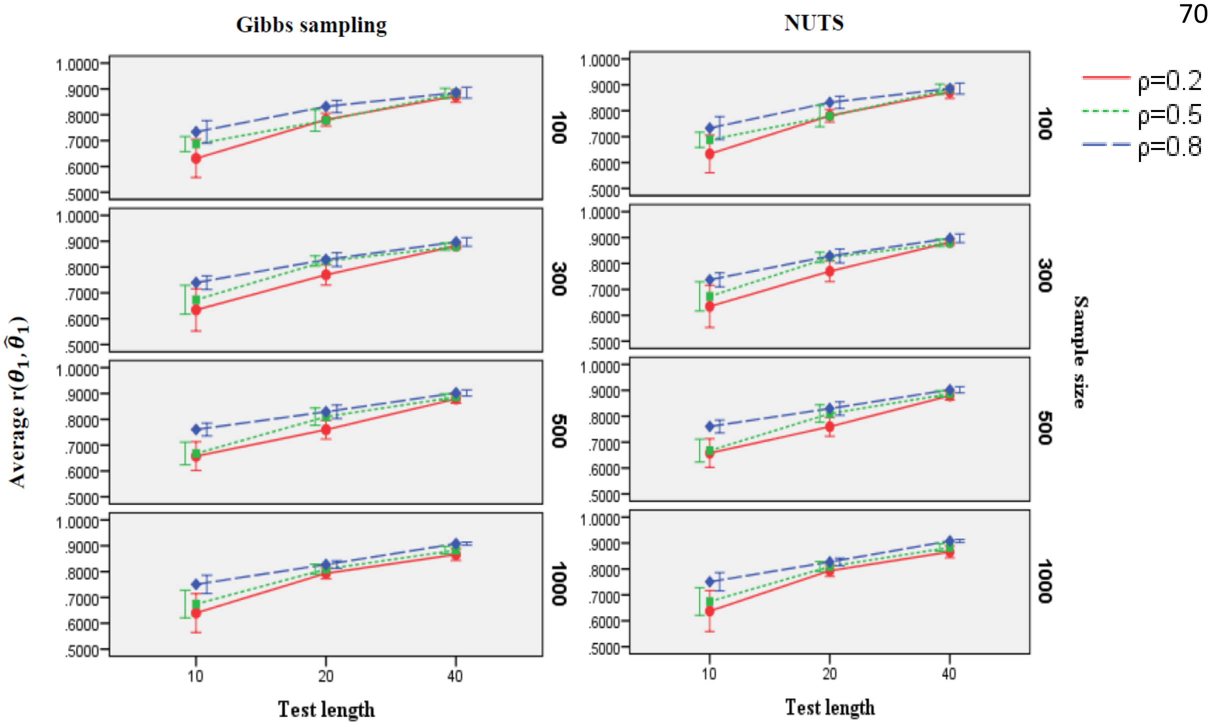

Figure 11. Average correlations between the actual and estimated person abilities $r(\theta_1,\hat{\theta}_1)$ under various test conditions for the 2PL multi-unidimensional model.

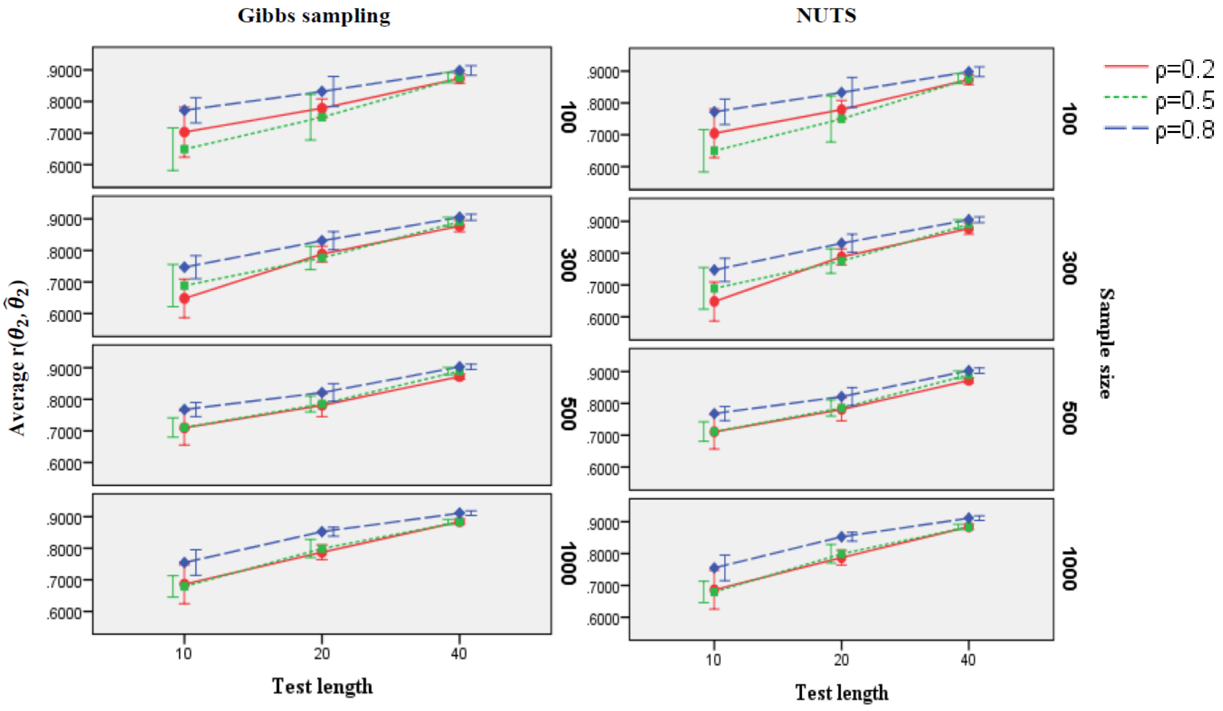

Figure 12. Average correlations between the actual and estimated person abilities $r(\theta_2,\hat{\theta}_2)$ under various test conditions for the 2PL multi-unidimensional model.

Table 11. Correlations between the actual and estimated person abilities $r(\theta_1,\hat{\theta}_1)$ and $r(\theta_2,\hat{\theta}_2)$ for the 2PL multi-unidimensional IRT model.

N	*K*		Gibbs sampling			NUTS		
			ρ_{12} = 0.2	ρ_{12} = 0.5	ρ_{12} = 0.8	ρ_{12} = 0.2	ρ_{12} = 0.5	ρ_{12} = 0.8
100	10	$r(\theta_1,\hat{\theta}_1)$	0.631	0.687	0.734	0.633	0.688	0.733
		$r(\theta_2,\hat{\theta}_2)$	0.703	0.649	0.772	0.705	0.650	0.772
	20	$r(\theta_1,\hat{\theta}_1)$	0.781	0.778	0.832	0.781	0.779	0.832
		$r(\theta_2,\hat{\theta}_2)$	0.779	0.750	0.832	0.779	0.750	0.833
	40	$r(\theta_1,\hat{\theta}_1)$	0.872	0.883	0.886	0.872	0.883	0.886
		$r(\theta_2,\hat{\theta}_2)$	0.873	0.875	0.898	0.873	0.876	0.898
300	10	$r(\theta_1,\hat{\theta}_1)$	0.634	0.673	0.740	0.634	0.673	0.737
		$r(\theta_2,\hat{\theta}_2)$	0.648	0.689	0.747	0.648	0.690	0.747
	20	$r(\theta_1,\hat{\theta}_1)$	0.770	0.823	0.829	0.770	0.824	0.829
		$r(\theta_2,\hat{\theta}_2)$	0.788	0.776	0.830	0.788	0.775	0.831
	40	$r(\theta_1,\hat{\theta}_1)$	0.882	0.879	0.897	0.882	0.879	0.897
		$r(\theta_2,\hat{\theta}_2)$	0.877	0.891	0.905	0.877	0.891	0.905
500	10	$r(\theta_1,\hat{\theta}_1)$	0.657	0.668	0.761	0.658	0.667	0.760
		$r(\theta_2,\hat{\theta}_2)$	0.710	0.711	0.768	0.710	0.711	0.768
	20	$r(\theta_1,\hat{\theta}_1)$	0.760	0.810	0.829	0.760	0.810	0.830
		$r(\theta_2,\hat{\theta}_2)$	0.781	0.785	0.821	0.781	0.786	0.821
	40	$r(\theta_1,\hat{\theta}_1)$	0.879	0.886	0.902	0.879	0.886	0.902
		$r(\theta_2,\hat{\theta}_2)$	0.872	0.890	0.903	0.872	0.890	0.903
1000	10	$r(\theta_1,\hat{\theta}_1)$	0.640	0.674	0.751	0.638	0.675	0.751
		$r(\theta_2,\hat{\theta}_2)$	0.686	0.679	0.755	0.686	0.680	0.755
	20	$r(\theta_1,\hat{\theta}_1)$	0.794	0.810	0.827	0.793	0.810	0.827
		$r(\theta_2,\hat{\theta}_2)$	0.787	0.799	0.853	0.787	0.799	0.853
	40	$r(\theta_1,\hat{\theta}_1)$	0.866	0.883	0.908	0.866	0.883	0.908
		$r(\theta_2,\hat{\theta}_2)$	0.884	0.883	0.911	0.884	0.883	0.911

4.2.3 Analysis of Variance (ANOVA) Results for the 2PL Multi-unidimensional Model

Effect sizes of the four factors, namely, sample size (*N*), test length (*K*), intertrait correlation (ρ_{12}), and MCMC algorithm (*A*), in the accuracy of estimating the 2PL multi-unidimensional IRT model under various simulated conditions are summarized in Table 12. Also, the interaction plots are presented in Figures 13 and 14.

For the discrimination parameters estimation, Table 12 shows that sample size (*N*) has the largest effect, accounting for 24.1% (26.7%) of the variance in the log*RMSE* (log*MAE*) of the discrimination parameter a_1 estimates. Test length (*K*) has a medium effect, accounting for 11.7% (11.5%) of the variance in the log*RMSE* (log*MAE*) of a_1 estimates. With a small effect size, intertrait correlation (ρ) contributes 2.1% (2.5%) of variance in the log*RMSE* (log*MAE*). The main effect of algorithm (*A*) and all interactions account for less than 1% of the variance in the log*RMSE* and log*MAE* of the discrimination parameter a_1 estimates. For example, the interaction between sample size (*N*) and test length (*K*) contributes about 0.8% of variance in the log*RMSE* of the discrimination parameter a_1 estimates. Similarly, it shows that sample size (*N*) has the largest effect, accounting for 31.4% (30.9%) of the variance in the log*RMSE* (log*MAE*) of the discrimination parameter a_2 estimates. Test length (*K*) has a medium effect, accounting for 11.6% (11.1%) of the variance in the log*RMSE* (log*MAE*) of a_2 estimates. In addition, the interaction between sample size (*N*) and intertrait correlation (ρ) has a small effect, contributing about 1.1% (0.9%) of the variance in the log*RMSE* (log*MAE*) of the discrimination parameter a_2 estimates. The main effect of intertrait correlation (ρ) and algorithm (*A*) together with other interactions contribute less than 1% of the variance in the log*RMSE* and log*MAE* of the discrimination parameter a_2 estimates. The interaction plot also suggests that there is no noticeable effect among these factors in estimating the discrimination parameters (a_1, a_2) (see Figure13).

For the difficulty parameters estimation, it shows that sample size (*N*) again has the largest effect, accounting for 12.9% (18.5%) of the variance in the log*RMSE* (log*MAE*) of the difficulty parameter b_1 estimates. Test length (*K*) has a small effect, accounting for 1.4% (2.3%) of the variance in the log*RMSE* (log*MAE*) of b_1 estimates. In addition, the interaction among sample size (*N*), test length (*K*), and intertrait correlation (ρ) has a small effect, contributing about 1.3% (0.8%) of the variance in the log*RMSE* (log*MAE*) of the difficulty parameter b_1 estimates, and the interaction between test length (*K*) and sample size (*N*) also has a small effect, accounting for 1.2% (1.5%) of the variance in the log*RMSE* (log*MAE*) of b_1 estimates. The other two main effects of intertrait correlation (ρ) and algorithm (*A*), and interactions account for no more than 1% of the variance in the log*RMSE* and log*MAE* in estimating b_1. Similarly, it shows that with a large effect size, sample size (*N*) accounts for 18.6% (22.6%) of the variance in the log*RMSE* (log*MAE*) of the difficulty parameter b_2 estimates. On the other hand, test length (*K*) only has a small effect, accounting for 2.3% (3.4%) of the variance in the log*RMSE* (log*MAE*) of b_2 estimates. In addition, the interaction among sample size (*N*), test length (*K*), and Intertrait correlation (ρ) has a small effect, contributing about 1.0% of variance in the log*MAE* of the difficulty parameter b_2 estimates. The other two main effects of intertrait correlation (ρ) and algorithm (*A*), and interactions account for less than 1% of the variance in the log*RMSE* and log*MAE* in estimating b_2. Similarly, the interaction plot suggests that there is no noticeable effect among these factors in estimating the difficulty parameters (b_1, b_2) (see Figure14).

For the person ability parameters estimation, it shows that test length (*K*) accounts for the majority of the variance, about 67.0% of the variance in the correlation between θ_1 and $\hat{\theta}_1$. Also, intertrait correlation (ρ_{12}) has a medium effect, contributing about 6.3% of the variance in the correlation between θ_1 and $\hat{\theta}_1$. In addition, interaction between test length (*K*) and intertrait correlation (ρ_{12}) has a small effect, accounting for 1.9% of the variance in the correlation between θ_1 and $\hat{\theta}_1$. The other two main effects and interactions contribute less than 1% of the variance in the correlation, with the maximum value of $\widehat{\omega}^2$ equal to 0.002, which is the interaction among test length (*K*), sample size (*K*) and intertrait correlation (ρ_{12}). In addition, test length (*K*) again accounts for most of the variance, about 61.5% of the variance in the correlation between θ_2 and $\hat{\theta}_2$. Similarly, intertrait correlation (ρ_{12}) has a medium effect, contributing about 6.6% of the variance in the correlation between θ_2 and $\hat{\theta}_2$. Moreover, interaction between test length (*K*) and intertrait correlation (ρ_{12}) has a small effect, accounting for 1.1% of the variance in the correlation between θ_2 and $\hat{\theta}_2$. The other two main effects and interactions, however, contribute less than 1% of the variance in the correlation, with the maximum value of $\widehat{\omega}^2$ equal to 0.004, which is the interaction between sample size (*N*) and intertrait correlation (ρ_{12}).

In summary, the ANOVA results support the conclusions drawn from Tables 7 through 11. Sample size plays a more important role than test length on the recovery of the discrimination parameters, with larger *N* leading to a better estimation. Also, sample size accounts for more proportion of variance in both log*RMSE* and log*MAE* of item parameters estimates from dimension 2 than those from dimension 1. On the other hand, intertrait correlation accounts for more proportion of variance in both log*RMSE* and log*MAE* of the discrimination parameter estimates from dimension 1 than those from dimension 2. In addition, test length and intertrait correlation positively affect the estimation of the person ability parameter, with more items and higher intertrait correlation leading to a better estimation. Also, test length accounts for more proportion of variance in the correlation between θ_1 and $\hat{\theta}_1$ than those in the correlations between θ_2 and $\hat{\theta}_2$.

Table 12. ANOVA effect sizes ($\widehat{\omega}^2$) for log*RMSE* in estimating the discrimination ($a_{1,}a_2$), difficulty ($b_{1,}b_2$) parameters, $r(\theta_1,\hat{\theta}_1)$, and $r(\theta_2,\hat{\theta}_2)$ in the 2PL multi-unidimensional IRT model.

Variable	log$RMSEa_1(a_2)$	log$MAEa_1(a_2)$	log$RMSEb_1(b_2)$	log$MAEb_1(b_2)$	$r(\theta_1,\hat{\theta}_1)$	$r(\theta_2,\hat{\theta}_2)$
K	0.117(0.116)	0.115(0.111)	0.014(0.023)	0.023(0.034)	0.670	0.616
N	0.241(0.314)	0.267(0.309)	0.129(0.186)	0.185(0.226)	0.000	0.002
A	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000	0.000
ρ	0.021(0.006)	0.025(0.003)	0.000(0.000)	0.000(0.000)	0.063	0.066
K×N	0.008(0.004)	0.007(0.005)	0.012(0.000)	0.015(0.004)	0.000	0.004
K×A	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000	0.000
K×ρ	0.006(0.003)	0.007(0.003)	0.004(0.009)	0.004(0.008)	0.019	0.011
N×A	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000	0.000
N×ρ	0.007(0.011)	0.006(0.009)	0.005(0.001)	0.010(0.001)	0.000	0.004
A×ρ	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000	0.000
K×N×A	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000	0.000
K×N×ρ	0.008(0.010)	0.008(0.007)	0.013(0.006)	0.008(0.010)	0.002	0.000
K×A×ρ	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000	0.000
N×A×ρ	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000	0.000
K×N×A×ρ	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000	0.000

Note. *K*: Test length; *N*: Sample size; *A*: Algorithm; *ρ*: Intertrait correlation.

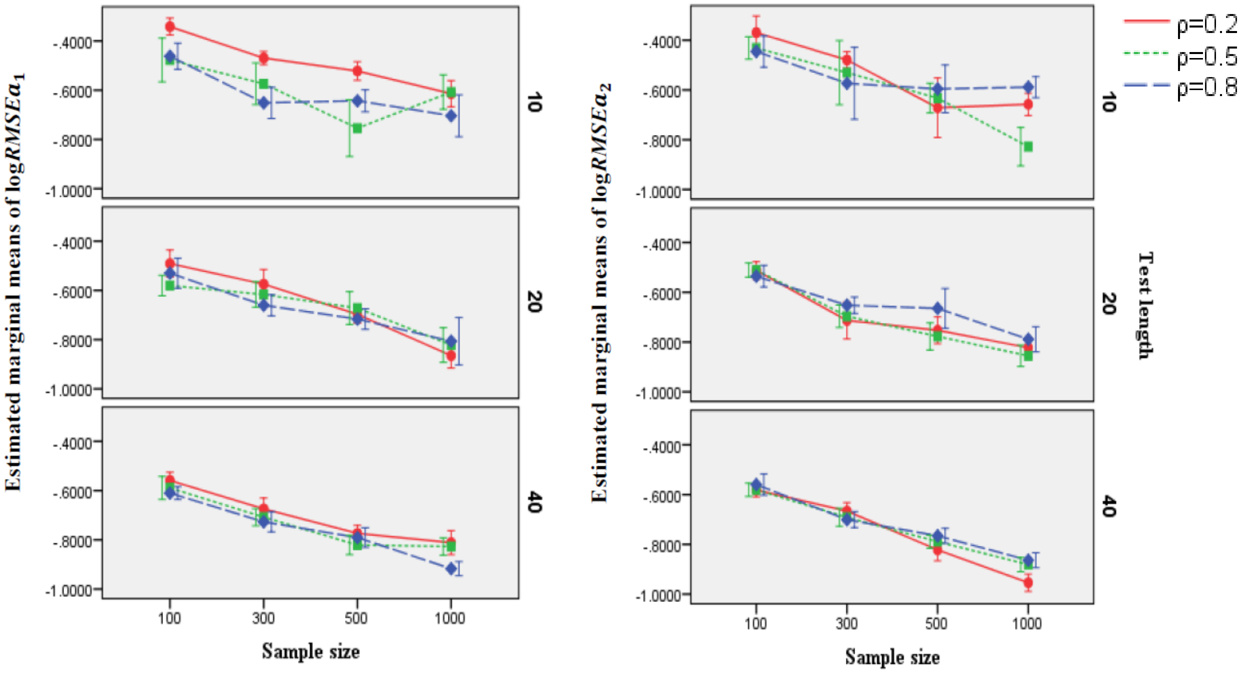

Figure 13. The interactions of sample size and intertrait correlation under different test lengths in logRMSEa_1 (left) and logRMSEa_2 (right).

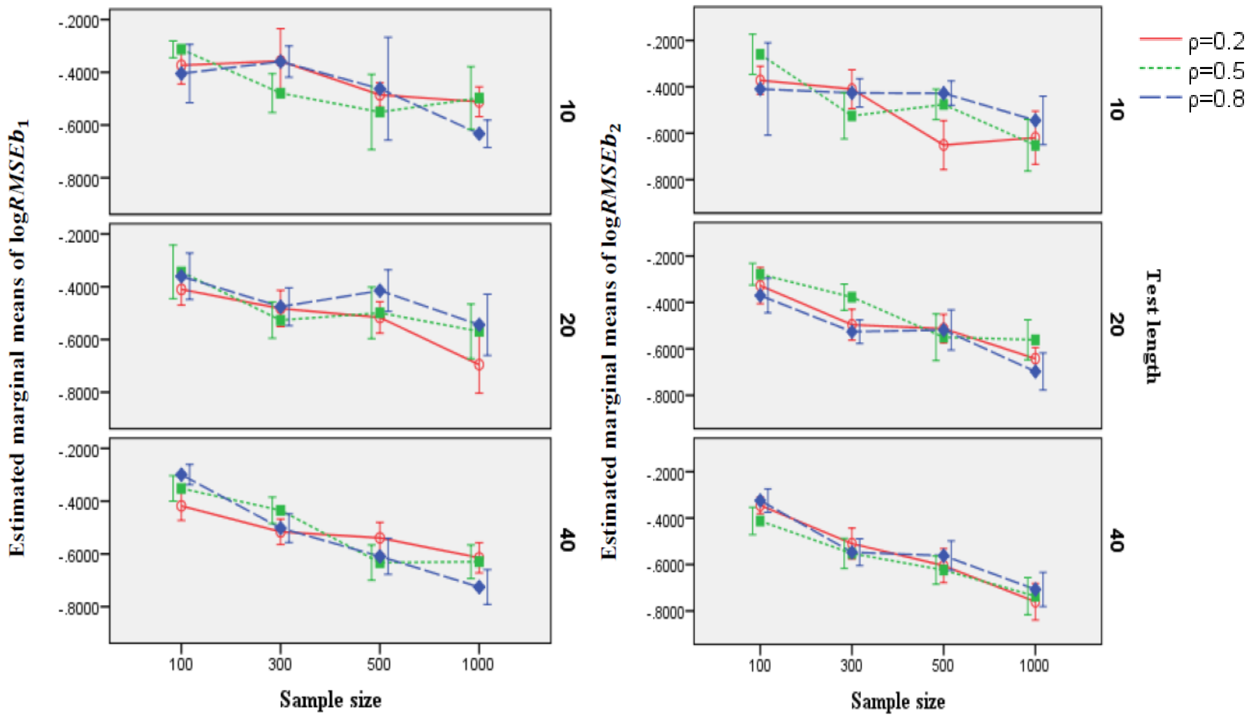

Figure 14. The interactions of sample size and intertrait correlation under different test lengths in logRMSEb_1 (left) and logRMSEb_2 (right).

CHAPTER 5

DISCUSSION AND CONCLUSION

This chapter contains two main sections. The first section summarizes the results on comparison of Gibbs sampling and No-U-Turn sampler (NUTS) for recovering parameters in the two-parameter logistic (2PL) unidimensional IRT (UIRT) model and the 2PL multi-unidimensional IRT model. Implications are discussed with the use of the Markov chain Monte Carlo (MCMC) algorithms for a fully Bayesian estimation when sample sizes, test lengths, prior specifications for a_j, and/or intertrait correlations vary. Then, Section 2 illustrates limitations of this dissertation and provides directions for future studies.

5.1 Comparison of Gibbs Sampling and NUTS for the 2PL IRT Models

The findings for the simulation studies are summarized and discussed in the following sections for the 2PL UIRT and 2PL multi-unidimensional IRT models.

5.1.1 Model Parameters Recovery for the 2PL UIRT Model

Simulation study 1 compares Gibbs sampling with NUTS in the performance of parameter recovery for the 2PL UIRT model via manipulating three factors: prior specifications for the item discrimination parameter (a_j), sample sizes (*N*), and test lengths (*K*). Results of the simulation study presented in Chapter 4 indicate that when comparing Gibbs sampling with NUTS, both fully Bayesian algorithms recover item and person ability parameters with similar accuracy and bias. More specifically, the two algorithms result in identical estimates under most conditions except for the condition with a lognormal prior for a_j and a small sample size (i.e., N=100). This further suggests that if the lognormal prior distribution is used for a_j in the 2PL UIRT model with sample sizes such as $N \leq 100$, NUTS rather than Gibbs sampling should be considered in order to obtain a better estimation of the discrimination parameters. In addition, the ANOVA results based on $\widehat{\omega}^2$ values also suggest that there is an interaction effect between prior specifications for a_j and sample size in estimating the discrimination parameters. For example, among the three prior distributions considered for a_j, the truncated normal prior should be adopted for a_j with either Gibbs sampling or NUTS when estimating the discrimination parameters with sample sizes such as $N \leq 300$. As the sample size increases (i.e., $N > 300$), the advantage of using the truncated normal prior for a_j is not that noticeable since the root mean square error (*RMSE*) and the mean absolute error (*MAE*) values of these three priors are similar. Theoretical explanation for the use of a lognormal prior or a truncated normal prior for a_j comes from the fact that in typical test settings, the a_j are assumed to be greater than zero, suggesting that the distribution of a_j can be specified as a unimodal and positively skewed distribution such as the lognormal (Mislevy, 1986) or truncated normal. In addition, a possible reason for the advantage of using the truncated normal prior distribution over lognormal prior distribution in estimating the discrimination parameters when sample size is small might be that when the range for the discrimination parameter a_j is from 0 to 0.5, the truncated normal prior distribution is more informative than the lognormal prior distribution (see Figure 15).

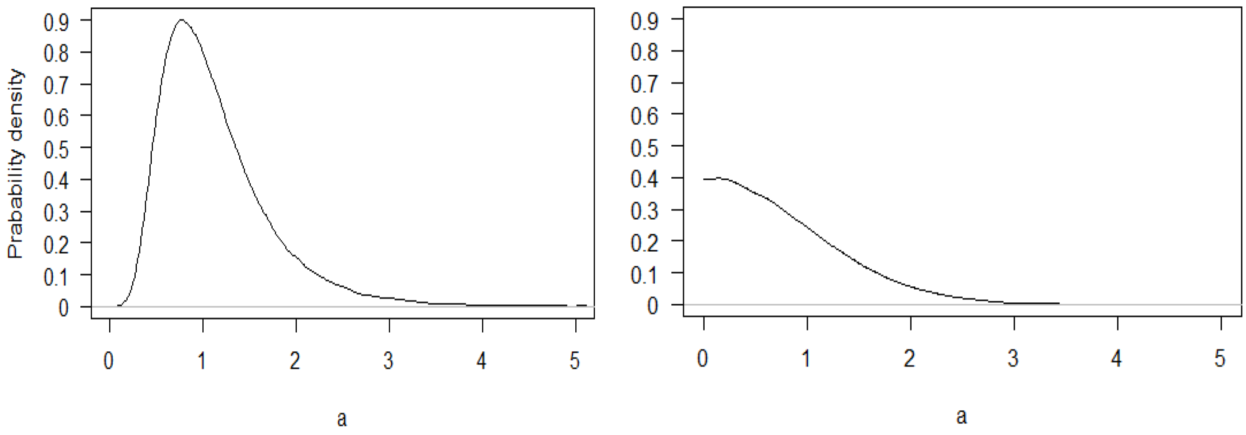

Figure 15. The probability density plot of lognormal(0, 0.5) distribution (left) and truncated normal $N_{(0,\infty)}$ (0, 1) distribution (right).

The results also indicate that sample size plays an important role in item parameter estimation. That is, increased sample sizes improve the precision in estimating both the discrimination and difficulty parameters. Since increased sample sizes provide more information on estimating items, item parameter estimation improves accordingly. Similarly, test length plays a more important role than other factors in improving the precision of the person ability parameter estimates. In other words, increased test lengths provide more information on subjects and hence, the person ability parameter can be better recovered. In terms of reducing the bias, increased sample sizes and test lengths have mixed effects. These findings are consistent with the previous studies, suggesting that sample size affects the accuracy of item parameter estimation and test length affects the accuracy of person ability parameter estimation (e.g., Kieftenbeld & Natesan, 2012; Roberts & Thompson, 2011; Sheng, 2010; Swamnathan & Gifford, 1982; Wollack, Bolt, Cohen, & Lee, 2002). It is, however, noted that test length only has a positive effect in estimating the discrimination parameters when data involve large sample sizes (i.e., $N \geq 300$), which is not consistent with previous research (e.g., Sheng, 2010).

In addition, when comparing the average *RMSE*s and *MAE*s under the condition where the prior distribution for a_j is lognormal in the UIRT model, there are some inconsistent results between the two algorithms mainly due to the reason that there is a slight difference in the estimated values. The *RMSE* and the *MAE* are considered as two of the most commonly used metrics for measuring the precision of parameter estimates in IRT models. Both metrics express average model prediction error in units of the variable of interest with smaller values suggesting a better accuracy. The *RMSE*, however, gives a relatively high weight to large errors since the errors are squared before they are averaged. In other words, the *RMSE* should be more useful when large errors are particularly undesirable. Also, the *RMSE* is more preferable to use than the

MAE when model errors follow a normal distribution (Chai & Draxler, 2014). From an interpretation perspective, the *MAE* is easier to understand than the *RMSE*. On the other hand, one distinct advantage of the *RMSE* over the *MAE* is that the *RMSE* avoids the use of taking the absolute value, which is undesirable in many mathematical calculations (Chai & Draxler, 2014). Moreover, Pelánek (2015) suggests that for the binary outcomes, the *MAE* metric should not be used since it is not a proper score and can lead to misleading conclusions.

5.1.2 Model Parameters Recovery for the 2PL Multi-unidimensional IRT Model

Simulation study 2 compares Gibbs sampling with NUTS in the performance of parameter recovery for the 2PL multi-unidimensional model via manipulating three factors: intertrait correlations (ρ_{12}), sample sizes (*N*), and test lengths (*K*). Again, when considering the effect of algorithms, the results on parameter recovery of the 2PL multi-unidimensional model indicate that Gibbs sampling and NUTS perform similarly across all conditions. In addition, increased sample sizes improve the precision but not the bias in estimating the discrimination parameter estimates, which is inconsistent with the results of the 2PL UIRT model in which as sample size increases, the average *RMSE*s/*MAE*s tend to decrease for both the discrimination and difficulty parameters. Test length, however, has no consistent effect on the accuracy or reducing bias in estimating item parameters. On the other hand, for the recovery of the person ability parameters (θ_1 and θ_2), the results suggest that test length and intertrait correlation have a positive and major effect on estimating θ_1 and θ_2. As discussed in Section 5.1.1, increased test lengths provide more information on subjects and therefore, the person ability parameter can be better recovered.

In addition, increased ρ_{12} suggests that the two latent traits have more overlap since one trait is closely related to the other. That is, a higher ρ_{12} enables the latent traits to share more information with one another and hence, the person ability parameters can be better recovered with relatively less information (i.e., fewer number of items). However, if ρ_{12} is low, the overall test is similar to measuring two separate sets of θs and therefore, more information is required in order to achieve a similar level of precision in estimation. In other words, more items are needed in order to better estimate the two separate sets of θs.

Also, the ANOVA results based on $\widehat{\omega}^2$ values indicate that there is an interaction effect between test length and intertrait correlation. This indicates that the effect of test length is not consistent across the three levels of intertrait correlations. Specifically, increased test lengths have a more positive influence on the accuracy of estimating person ability parameters when there is a low intertrait correlation (e.g., ρ_{12}=0.2) than when there is a high intertrait correlation (e.g., ρ_{12}=0.8). In addition, the values of $r(\theta_1,\hat{\theta}_1)$ and $r(\theta_2,\hat{\theta}_2)$ are similar among the three levels of intertrait correlations when K=40. However, the values of $r(\theta_1,\hat{\theta}_1)$ and $r(\theta_2,\hat{\theta}_2)$ are higher for ρ_{12}=0.8 than those for ρ_{12}=0.2 when K=10. It is noted that although higher intertrait correlation suggests a better recovery of θ_1 and θ_2, in real test situations, ρ_{12} is not readily known and needs to be estimated. From the results presented in Chapter 4, the correlations between the true and estimated person ability parameters $r(\theta_1,\hat{\theta}_1)$ and $r(\theta_2,\hat{\theta}_2)$ are all about 0.9 when K=40. Therefore, when there are a sufficient number of items, both algorithms can obtain equally accurate estimates of θ_1 and θ_2 regardless of ρ_{12}. This also implies that even though tests with less correlated dimensions do not estimate θ_1 and θ_2 as well as those with highly correlated dimensions, one can compensate it by increasing test length.

In terms of recovering the intertrait correlation (ρ_{12}), the results suggest that as ρ_{12} increases, the error and bias for estimating the intertrait correlation increase for $N\geq300$. Also, all negative biases suggest that the true parameter is consistently underestimated. One possible explanation of this result is the use of the inverse-Wishart prior for the covariance matrix in the present study. Alvarez, Niemi, and Simpson (2014) pointed out that an inverse-Wishart prior might not work well when the true variance is small relative to the prior mean under which the posterior for the variance is biased toward larger values and the correlation is biased toward zero. This bias remains even for data with large sample sizes and therefore, caution should be used when using the inverse-Wishart prior. The other issues of using this prior, which can impact posterior inferences about the covariance matrix include that the uncertainty for all variances is set by a single degree of freedom parameter (Gelman, 2014), the marginal distribution for the variances has low density in a region near zero (Gelman, 2006), and there is an *a priori* dependence between correlations and variances (Tokuda, Goodrich, Van Mechelen, Gelman, & Tuerlinckx, 2011). Due to these reasons, the Stan manual (Stan Development Team, 2016) suggests the LKJ prior for the correlation matrices (Lewandowski, Kurowicka, & Joe, 2009). From a modeling perspective, the LKJ prior is appealing since one can model correlations and variances independently and allow the data to define their relationships. The main disadvantage of the LKJ prior, however, is computational. Currently, the LKJ prior cannot be used in the programs such as JAGS based on Gibbs sampling.

5.1.3 Computational Speed of Gibbs Sampling and NUTS

In terms of computational speed of the two algorithms implemented in the two programs utilizing computers with a processor 2.7 GHz Intel Core i5 and memory 8 GB 1600 MHz, under exactly the same condition, Gibbs sampling takes a longer computational time than NUTS to fit the 2PL UIRT model. For example, the computation time of implementing Gibbs sampling in JAGS to data with N=1000 and K=20 was about 73 minutes to complete four chains with 5000 iterations. For the same data size and number of iterations, NUTS via the use of Stan took about

32 minutes. This is in line with findings from the previous study that NUTS is more efficient than Gibbs sampling (Grant et al., 2016). However, when it comes to the 2PL multi-unidimensional IRT model, NUTS takes a longer time than Gibbs sampling. For example, the computation time of implementing Gibbs sampling in JAGS to data with N=1000 and K=40 was about 38 minutes to complete four chains with 5000 iterations. For the same data size and number of iterations, NUTS via the use of Stan took about 141 minutes. This is obviously different from findings on the 2PL UIRT model or from Grant et al. (2016). A possible cause might be that the 2PL multi-unidimensional model is more complicated than the 2PL UIRT model or the Rasch model considered by Grant et al. (2016). Moreover, the Stan code used in the present study for the 2PL multi-unidimensional model has a slow mixing issue for some iterations and needs further modification to make it more efficient. Further investigations are needed to understand the actual reason. In addition, when comparing the computation time of the two algorithms, the speed difference may be due to the use of two different software packages. It will be ideal that the two procedures are compared in the same setting. This is, however, difficult due to availability of MCMC software. Readers need to take this into consideration when interpreting results of the present study.

5.2 Limitations and Directions for Future Studies

Through simulation studies, this dissertation shows that researchers and practitioners should benefit from using Gibbs sampling and NUTS in estimating parameters of the 2PL UIRT and 2PL multi-unidimensional IRT models. More importantly, the results of the present study show that Gibbs sampling and NUTS perform equally well across most of the simulated conditions. There is not much difference in the accuracy/bias using Gibbs sampling vs. NUTS when implementing them to the 2PL IRT models. The results also provide some sense of assurance that decisions about which algorithm to use should be considered other than accuracy in estimation. It is, however, noted that conclusions are based on simulated conditions considered in the present study and cannot be generalized to other conditions. For example, the present study only considers four sample sizes (i.e., 100, 300, 500, and 1000 examinees), three test lengths (i.e., 10, 20, and 40 items), three intertrait correlations (i.e., 0.2, 0.5, and 0.8), and equal test items between two latent traits for the 2PL multi-unidimensional model, but for future studies, additional test conditions need to be explored too. In addition, the results of this study are based on up to 25 replications due to the fact that MCMC algorithms are computationally expensive taking considerable time to execute, making it difficult to go with 25 replications in this dissertation for all simulated conditions. Given the small number of iterations, and given that Harwell et al. (1996) suggested a minimum of 25 replications for Monte Carlo studies in typical IRT-based research, *bias*, *RMSE*, and *MAE* values presented in Chapter 4 need to be verified with further studies before one can generalize the results to similar conditions.

In addition, the lack of the difference between Gibbs sampling and NUTS is likely due to the fact that the 2PL model is relatively simple and thus, does not have issues with mixing when it comes to implementing MCMC algorithms. Future studies can consider using a more complicated model such as the three-parameter logistic (3PL) model to compare the two algorithms since the 3PL model is a mixture model and requires more attention with convergence when compared with the 2PL model (Sheng, 2010).

Simulation studies often demonstrate performance under ideal situations. In this case, the true IRT model was known and fit can be assumed nearly perfect. Future studies may use these two MCMC algorithms to fit the 2PL IRT models to real data and use them for model comparison and selection. Other test format conditions should be also explored. For example, the present study only compares two MCMC algorithms and therefore, current simulation conditions could be expanded to compare other MCMC algorithms (e.g., Metropolis-Hastings and Hastings-within-Gibbs). Also, future studies can compare the fully Bayesian estimation with other estimation methods such as the marginal maximum likelihood (MML) estimation (see Appendix A for a demonstration of the advantages of fully Bayesian estimation over MML). In addition, simulation study 2 only considers two latent dimensions. Future studies can compare the two algorithms on multi-unidimensional models that have more than two latent dimensions or the more general multidimensional IRT models. Moreover, the findings of the present study are limited to the dichotomous models. Models with polytomous categories (e.g., the partial credit or graded response models) should be also explored in future studies. There are a large number of choices for prior distributions or simulated values for the IRT model parameters. Due to that, other prior specifications or simulated values for model parameters a_j, b_j, and θ_i should also be considered. For example, future studies may use these two algorithms to fit IRT models with non-normal latent trait distributions (e.g., from a gamma distribution with a shape and scale parameter of 10 and 1.5). In addition, other priors such as the scaled inverse-Wishart, hierarchical inverse-Wishart, and LKJ priors should be explored for the covariance matrix for the multi-unidimensional IRT models. In addition to *RMSE* and *MAE*, other evaluation metrics such as area under the curve (AUC; Swets & Pickett, 1982) can be used in future studies as well.

REFERENCES

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-53. doi:10.1111/j.1745-3992.2003.tb00136.x

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied psychological measurement*, *21*(1), 1-23. doi: 10.1177/0146621697211001

Albert, J. H. (1992). A Bayesian analysis of a Poisson random effects model for home run hitters. *The American Statistician*, *46*(4), 246-253.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669-679.

Almond, R. G. (2014). A Comparison of Two MCMC Algorithms for Hierarchical Mixture Models. In *BMA@ UAI* (pp. 1-19).

Alvarez, I., Niemi, J., & Simpson, M. (2014). Bayesian inference for a covariance matrix. *arXiv preprint arXiv:1408.4050*.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological), 32*(2), 283-301.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*(1), 123-140. doi:10.1007/BF02291180

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement, 11*(2), 111-141. doi: 10.1177/014662168701100201

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement*, *22*(2), 153-169. doi: 10.1177/01466216980222005

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory : Parameter estimation techniques.* New York : Marcel Dekker. 2nd ed.

Barton, M. A., & Lord, F. M. (1981). An Upper Asymptote for the Three-Parameter Logistic Item-Response Model. *ETS Research Report Series*, *1981*(1), i-8. doi: 10.1002/j.2333-8504.1981.tb01255.x

Béguin, A. A., & Glas, C. A. W. (1998). *MCMC estimation of multidimensional IRT models*. Faculty of Educational Science and Technology, University of Twente.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*(4), 541-561. doi:10.1007/BF02296195

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology, 6*(2), 258-276. doi:10.1016/0022-2496(69)90005-4

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459. doi:10.1007/BF02293801

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261-280. doi:10.1177/014662168801200305

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. doi:10.1177/0146621603258350

Bradlow, E. T., Wainer, H., & Wang, X. H. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153-168. doi:10.1007/BF02294533

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434-455.

Caughey, D., & Warshaw, C. (2014). Dynamic Representation in the American States, 1960–2012. In *American Political Science Association 2014 Annual Meeting Paper*.

Chai, T., & Draxler, R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247-1250.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*(4), 327-335.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Copelovitch, M., Gandrud, C., & Hallerberg, M. (2015). Financial Regulatory Transparency, International Institutions, and Borrowing Costs. In *The Political Economy of International Organizations Annual Conference, University of Utah, Salt Lake City, Utah. http://wp. peio. me/wp-content/uploads/PEIO8/Copelovitch, Gandrud, Hallerberg* (Vol. 3, p. 2015).

Culpepper, S. A. (2015). Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika*, 1-22. doi:10.1007/s11336-015-9477-6

Dawber, T., Rogers, W. T., & Carbonaro, M. (2009). Robustness of Lord’s Formulas for Item Difficulty and Discrimination Conversions between Classical and Item Response Theory Models. *Alberta Journal of Educational Research, 55*(4), 512-533.

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement*, *16*(4), 327-343.

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

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*(3), 295-311. doi: 10.3102/10769986030003295

de la Torre, J., Stark, S., & Chernyshenko, O. S. (2006). Markov chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement, 30*(3), 216-232. doi: 10.1177/0146621605282772

DeMars, C. E. (2005). Scoring Subscales Using Multidimensional Item Response Theory Models. Poster presented at the annual meeting of the American Psychological Association, Washington, DC.

DeMars, C. E. (2010). *Item response theory*. Oxford ; New York : Oxford University Press.

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

Drasgow, F., & Schmitt, N. (2002). *Measuring and analyzing behavior in organizations : advances in measurement and data analysis*: San Francisco, CA : Jossey-Bass.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B, 195*, 216-222. doi:10.1016/0370-2693(87)91197-X

Eaves, L., Erkanli, A., Silberg, J., Angold, A., Maes, H. H., & Foley, D. (2005). Application of Bayesian Inference using Gibbs Sampling to Item-Response Theory Modeling of Multi-

Symptom Genetic Data. *Behavior Genetics, 35*(6), 765-780. doi:10.1007/s10519-005-7284-z

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(1), 5-18. doi:10.1007/s11136-007-9198-0

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381. doi:10.1177/0013164498058003001

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*(2), 271-288. doi:10.1007/BF02294839

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on an Article by Browne and Draper). *Bayesian Analysis, 1*(3), 515-533.

Gelman, A. (2014). *Bayesian data analysis*: Boca Raton : CRC Press. Third edition.

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics, 40*(5), 530-543. doi:10.3102/1076998615606113

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457-472.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*(6), 721-741. doi:10.1109/TPAMI.1984.4767596

Ghitza, Y., & Gelman, A. (2014). The great society, Reagan's revolution, and generations of presidential voting. *To be submitted*.

Ghosh, M. (1995). Inconsistent maximum likelihood estimators for the Rasch model. *Statistics & Probability Letters*, *23*(2), 165-170. doi:10.1016/0167-7152(94)00109-L

Ghosh, M., Ghosh, A., Chen, M. H., & Agresti, A. (2000). Noninformative priors for one-parameter item response models. *Journal of Statistical Planning and Inference*, *88*(1), 99-115.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*: London : Chapman & Hall.

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, *43*(1), 169-177. doi: 10.2307/2348941

Grant, R. L., Furr, D. C., Carpenter, B., & Gelman, A. (2016). Fitting Bayesian item response models in Stata and Stan. arXiv preprint arXiv:1601.03443.

Griewank, A., & Walther, A. (2008). *Evaluating derivatives: Principles and techniques of algorithmic differentiation*: Philadelphia, PA : Society for Industrial and Applied Mathematics. 2nd ed.

Gruijter, D. N. (1990). A note on the bias of UCON item parameter estimation in the Rasch model. *Journal of Educational Measurement*, *27*(3), 285-288. doi: 10.1111/j.1745-3984.1990.tb00749.x

Güler, N., Uyanık, G. K., & Teker, G. T. (2014). Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Research on Education*, *2*(1), 1-6.

Gulliksen, H. (1987). *Theory of mental tests*: Hillsdale, N.J. : L. Erlbaum Associates.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data1, 2, 3. *Journal of Educational Measurement*, *14*(2), 75-96. doi: 10.1111/j.1745-3984.1977.tb00030.x

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47. doi:10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*: Boston : Kluwer-Nijhoff Pub. ; Hingham, MA, U.S.A.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*: Newbury Park, Calif. : Sage Publications.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101-125. doi:10.1177/014662169602000201

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109. doi: 10.1093/biomet/57.1.97

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement*, *19*(4), 337-352. doi:10.1177/014662169501900404

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593-1623.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62*(2), 171-189. doi: 10.1007/BF02295273

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*: Hillsdale : Lawrence Erlbaum Associates.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of Two- and Three-Parameter Logistic Item Characteristic Curves: A Monte Carlo Study. *Applied Psychological Measurement, 6*(3), 249. doi:10.1177/014662168200600301

Huo, Y., de la Torre, J., Mun, E.-Y., Kim, S.-Y., Ray, A. E., Jiao, Y., & White, H. R. (2015). A Hierarchical Multi-Unidimensional IRT Approach for Analyzing Sparse, Multi-Group Data for Integrative Data Analysis. *Psychometrika, 80*(3), 834-855. doi:10.1007/s11336-014-9420-2

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25*(3), 285-306. doi: 10.3102/10769986025003285

Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software, 20*(10), 1-24.

Johnson, M. S., & Sinharay, S. (2005). Calibration of polytomous item families using Bayesian hierarchical modeling. *Applied Psychological Measurement, 29*(5), 369-400. doi: 10.1177/0146621605276675

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*(4), 331-358. doi: 10.1177/0146621606292213

Kieftenbeld, V., & Natesan, P. (2012). Recovery of Graded Response Model Parameters: A Comparison of Marginal Maximum Likelihood and Markov Chain Monte Carlo Estimation. *Applied Psychological Measurement, 36*(5), 399-419. doi: 10.1177/0146621612446170

Kim, S. H. (2007). Some posterior standard deviations in item response theory. *Educational and Psychological Measurement*, *67*(2), 258-279. doi: 10.1177/0013164407067002050 1

Knol, D. L., & Berger, M. P. (1991). Empirical Comparison Between Factor Analysis and Multidimensional Item Response Models. *Multivariate Behavioral Research*, *26*(3), 457-477. doi:10.1207/s15327906mbr2603_5

Lee, H. (1995). *Markov chain Monte Carlo methods for estimating multidimensional ability in item response theory* (Unpublished doctoral dissertation). University of Missouri, Columbia.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989-2001.

Lim, R. G., & Drasgow, F. (1990). Evaluation of Two Methods for Estimating Item Response Theory Parameters When Assessing Differential Item Functioning. *Journal of Applied Psychology*, *75*(2), 164-174.

Linden, W. J. v. d., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*: Dordrecht ; Boston : Kluwer Academic.

Lord, F. M. (1952). *A theory of test scores*: New York, Psychometric Society.

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

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

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*(4), 325-337. doi:10.1023/A:1008929526011

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. doi:10.1007/BF02296272

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087-1092.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*(247), 335-341.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177-195. doi:10.1007/BF02293979

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176. doi:10.1002/j.2333-8504.1992.tb01436.x

Neal, R. M. (1992). An improved acceptance procedure for the hybrid Monte Carlo algorithm. *arXiv preprint hep-lat/9208011*.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, *2*, 113-162.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, *16*(1), 1-32. doi: 10.2307/1914288

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and Demonstration of Multidimensional IRT-Based Internal Measures of Differential Functioning of ltems and Tests. *Journal of Educational Measurement*, *34*(3), 253-272. doi:10.1111/j.1745-3984.1997.tb00518.x

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146-178. doi: 10.3102/10769986024002146

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342-366. doi: 10.3102/10769986024004342

Pelánek, R. (2015). Metrics for Evaluation of Student Models. *Journal of Educational Data Mining*, *7*(2), 1-19.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, p. 125).

Plummer, M. (2013). rjags: Bayesian graphical models using MCMC. *R package version*, *3*.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*: Copenhagen.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401-412. doi:10.1177/014662168500900409

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36. doi: 10.1177/0146621697211002

Reckase, M. D. (2009). *Multidimensional item response theory*. New York; London : Springer.

Rijmen, F. (2009). Efficient full information maximum likelihood estimation for multidimensional IRT models. *ETS Research Report Series*, *2009*(1), i-31. doi: 10.1002/j.2333-8504.2009.tb02160.x

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1-25. doi: 10.18637/jss.v017.i05

Roberts, J., & Thompson, V. (2011). Marginal Maximum A Posteriori Item Parameter Estimation for the Generalized Graded Unfolding Model. *Applied Psychological Measurement*, *35*(4), 259-279. doi: 10.1177/0146621610392565

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation*, *72*(3), 217-232.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. No. 17. Richmond, VA: Psychometric Society.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph Supplement*. No. 18. Richmond, VA: Psychometric Society.

Sheng, Y. (2008). A MATLAB package for Markov chain Monte Carlo with a multi-unidimensional IRT model. *Journal of Statistical Software*, *28,* 1-20.

Sheng, Y. (2010). A sensitivity analysis of Gibbs sampling for 3PNO IRT models: Effects of prior specifications on parameter estimates. *Behaviormetrika*, *37*(2), 87-110. doi: 10.2333/bhmk.37.87

Sheng, Y., & Headrick, T. C. (2012). A Gibbs Sampler for the Multidimensional Item Response Model. *ISRN Applied Mathematics*. Article 269385, 1–14. doi:10.5402/2012/269385

Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, *67*(6), 899-919. doi:10.1177/0013164406296977

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement, 68*(3), 413-430. doi: 10.1177/0013164407308512

Sheng, Y., & Wikle, C. K. (2009). Bayesian IRT models incorporating general and specific abilities. *Behaviormetrika*, *36*(1), 27-48.

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

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research*, *56*(4), 495-529. doi:10.3102/00346543056004495

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS Version 1.4 User Manual.* MRC Biostatistics Unit. URL http://www.mrc-bsu.cam.ac.uk/bugs/.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D. (2010). *OpenBUGS Version 3.1.1 User Manual*. Helsinki, Finland.

Stan Development Team. (2016). *Stan Modeling Language Users Guide and Reference Manual*, Version 2.12.0.

Su, Y. S., & Yajima, M. (2012). R2jags: A Package for Running jags from R. *R package version 0.03-08*.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian Estimation in the Rasch Model. *Journal of Educational Statistics*, *7*(3), 175-191. doi: 10.2307/1164643

Swaminathan, H., & Gifford, J. A. (1983). Estimation of Parameters in the Three-Parameter Latent Trait Model. *New Horizon Testing*, 13-30. doi:10.1016/B978-0-12-742780-5.50009-3

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems : Methods from signal detection theory*. New York : Academic Press.

Sympson, J. B. (1978). A model for testing with multidimensional items. In *Proceedings of the 1977 computerized adaptive testing conference* (pp. 82-98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions*: New York : Springer. 3rd ed.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528-540.

Thissen, D., & Wainer, H. (2001). *Test scoring*: Mahwah, N.J. : L. Erlbaum Associates, 2001.

Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006). Making BUGS open. *R news*, *6*(1), 12-17.

Tokuda, T., Goodrich, B., Van Mechelen, I., Gelman, A., & Tuerlinckx, F. (2011). Visualizing distributions of covariance matrices. *Unpublished Manuscript.*

Traub, R. E. (1997). Classical Test Theory in Historical Perspective. *Educational Measurement: Issues and Practice*, *16*(4), 8-14. doi:10.1111/j.1745-3992.1997.tb00603.x

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In *Computerized adaptive testing: Theory and practice* (pp. 245-269). Springer Netherlands. doi:10.1007/0-306-47531-6_13

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*: New York: Cambridge University Press.

Walker, C. M., & Beretvas, S. N. (2000). *Using Multidimensional versus Unidimensional Ability Estimates To Determine Student Proficiency in Mathematics*. In *American Educational Research Association 2000 Annual Meeting Paper*.

Wang, W. C., Wilson, M. R., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. *Objective Measurement: Theory into Practice*, *4*, 139-155.

Wang, X. H., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, *26*(1), 109-128. doi:10.1002/j.2333-8504.2002.tb01869.x

Whitely, S. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*(4), 479-494.

Wilson, D. T., Wood, R., & Gibbons, R. D. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis*. SSI, Scientific Software International.

Wingersky, M. S. (1992). Significant Improvements to LOGIST. *ETS Research Report Series*, *1992*(1), i-42. doi:10.1002/j.2333-8504.1992.tb01453.x

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *26*(3), 339-352. doi: 10.1177/0146621602026003007

Wright, B. D., & Stone, M. H. (1979). *Best test design : Rasch measurement*. Chicago : Mesa Press.

Yao, L., & Boughton, K. A. (2007). A Multidimensional Item Response Modeling Approach for Improving Subscale Proficiency Estimation and Classification. *Applied Psychological Measurement, 31*(2), 83-105. doi:10.1177/0146621606291559

Zhao, Y., & Hambleton, R. K. (2009). *Software for IRT analyses: Descriptions and features*. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Zheng, B. (2000). Bayesian estimation of multidimensional item response theory model using Gibbs sampling. *Communications in Statistics-Theory and Methods*, *29*(5-6), 1405-1417.

Zhu, L., Robinson, S. E., & Torenvlied, R. (2014). A Bayesian Approach to Measurement Bias in Networking Studies. *American Review of Public Administration*, *45*(5), 542-564. doi: 10.1177/0275074014524299

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

APPENDICES

APPENDIX A

Comparing MML and Fully Bayesian Estimations with An IRT Model

In order to demonstrate the advantages of the fully Bayesian estimation over the conventional marginal maximum likelihood (MML) estimation, a few simulations were carried out. As described in Chapter 2, Bayesian estimation has advantages over MML when data involve small samples and short tests or when an unusual response pattern occurs. To empirically demonstrate this, data were simulated from the two-parameter logistic (2PL) unidimensional IRT (UIRT) model as defined in Equation (3.2.1). For the small sample and short test condition, sample size (N) was manipulated to be 10, 50, and 100 examinees and test length (K) was fixed at 10 items. Model parameters were generated such that $\theta_i \sim N(0, 1)$, $a_j \sim U(0, 2)$, and $b_j \sim U(-2, 2)$. For the all correct/incorrect response pattern condition, sample size was set to be 1000 examinees and test length to be 10 items. Model parameters were generated such that $\theta_i \sim N(0, 1)$, $a_j \sim U(0, 2)$, and either $b_j \sim U(-2, -1.9)$, suggesting all extremely easy items, or $b_j \sim U(1.9, 2)$, suggesting all extremely difficult items.

An R package ltm (Rizopoulos, 2006) was used for estimating the 2PL UIRT model using MML to compare with results from the fully Bayesian estimation using Gibbs sampling and No-U-Turn Sampler (NUTS). For the MCMC procedures, conjugate normal priors were assumed for both θ_i and b_j such that $\theta_i \sim N(0, 1)$, $b_j \sim N(0, 1)$, and $a_j \sim N_{(0,\infty)}(0, 1)$.

Gibbs sampling and NUTS were implemented to each simulated data via the use of JAGS and Stan, respectively, where the burn-in (or warm-up) stage was set to 3000 iterations followed by 4 chains with 5000 iterations. For both algorithms, the initial values for the discrimination parameters a_j were set to ones, and those for the difficulty parameters b_j and latent ability parameters θ_i were set to zeros. Further, convergence of Markov chains was evaluated using the Gelman-Rubin R statistic (Gelman & Rubin, 1992).

For each simulated condition, only one replication was conducted. The accuracy of item parameter estimates was evaluated using the average square error (*ASE*) and can be defined as:

$$ASE = \frac{\sum_{j=1}^{K}(\hat{\pi}_j - \pi_j)^2}{K} \tag{A.1}$$

where π is the true value of an item parameter, $\hat{\pi}$ is the estimated value of that parameter, and *K* is the total number of items.

The *ASE* measures the average of the squares of the errors between the true and estimated item parameters. In general, a small value of the *ASE* suggests a more accurate estimation of the item parameters.

The values of the true and estimated item parameters, and *ASE* values for each simulated condition by implementing MML, Gibbs sampling, or NUTS to recover the discrimination (a_j) and difficulty parameters (b_j) are presented in Tables A1 through A5.

The results show that both Gibbs sampling and NUTS perform equally well and better than MML under all simulated conditions. For example, the *ASE* values for the discrimination parameters using MML, Gibbs sampling, and NUTS are 433.167, 0.247, and 0.248, respectively with *N*=10 and *K*=10 (see Table A1). As sample size increases, the *ASE* values of either the discrimination or difficulty parameters using all three estimation methods decrease too. For example, as *N* increases from 10 to 100, the *ASE* values for the discrimination parameters using MML, Gibbs sampling, and NUTS decrease from 433.167, 0.247, and 0.248 to 0.554, 0.107, and 0.280, respectively (see Tables A1 and A3). In addition, for the all correct/incorrect response condition, the fully Bayesian estimation still outperforms MML even though the difference is not that noticeable as compared with the small sample and short test conditions. For example, the

ASE values for the discrimination parameters using MML, Gibbs sampling, and NUTS are 0.055, 0.045, and 0.045, respectively with *N*=1000 and extremely easy items (see Table A4).

As far as the computation time is concerned, the implementation of MML for estimating the 2PL UIRT model is very fast using the R package ltm. For example, with a processor 2.5 GHz Intel Core i5 and memory 8 GB 1600 MHz, the computation time of implementing Gibbs sampling in JAGS to data with *N*=100 and *K*=10 was about 55.41 seconds to complete four chains with 5000 iterations. For the same data size, MML via the use of ltm only took 0.11 seconds.

Overall, although computationally more expensive, the use of fully Bayesian estimation for the 2PL UIRT model results in a better recovery of item parameters compared to those via the use of the MML estimation for data with small sample sizes and short test lengths, or when all correct/incorrect response patterns occur.

Table A1. The true and estimated values of the item parameters in the 2PL UIRT model using MML, Gibbs sampling, and NUTS when *N*=10.

	True		MML		Gibbs sampling		NUTS	
Item	a	b	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
1	0.993	0.183	0.045	9.123	0.731	0.206	0.719	0.207
2	1.454	-1.962	-1.252	1.514	0.778	-0.887	0.766	-0.884
3	0.342	1.285	54.127	1.006	0.865	0.792	0.854	0.795
4	1.738	-0.546	-0.240	5.900	0.733	-0.888	0.731	-0.877
5	1.069	-1.369	1.287	0.049	0.619	-0.052	0.618	-0.055
6	0.344	1.214	-31.365	-0.072	0.375	0.174	0.373	0.174
7	0.397	1.439	1.114	1.025	0.787	0.487	0.767	0.480
8	0.591	1.036	0.178	12.434	1.081	1.185	1.083	1.182
9	0.769	-0.411	-0.020	0.042	0.546	-0.041	0.527	-0.041
10	0.858	1.242	21.350	0.684	0.746	0.473	0.739	0.481
MSE			433.167	26.789	0.247	0.599	0.248	0.598

Table A2. The true and estimated values of the item parameters in the 2PL UIRT model using MML, Gibbs sampling, and NUTS when *N*=50.

	True		MML		Gibbs sampling		NUTS	
Item	*a*	*b*	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
1	1.091	-1.755	0.085	-17.887	0.741	-1.849	0.730	-1.873
2	1.015	0.755	1.768	0.368	1.209	0.488	1.204	0.484
3	1.380	-1.344	2.183	-1.144	1.568	-1.337	1.553	-1.350
4	1.463	-0.151	1.243	0.017	1.006	0.036	0.993	0.047
5	0.895	-0.005	2.143	-0.118	1.485	-0.114	1.472	-0.119
6	0.880	1.512	1.706	0.931	1.360	1.099	1.354	1.102
7	0.573	0.334	1.216	-0.325	0.821	-0.385	0.818	-0.383
8	1.732	-1.509	6.665	-1.056	1.843	-1.465	1.842	-1.472
9	1.318	-0.015	1.320	0.098	1.093	0.138	1.083	0.131
10	1.213	0.737	0.731	1.169	0.733	1.056	0.731	1.048
MSE			2.949	26.165	0.134	0.094	0.133	0.094

Table A3. The true and estimated values of the item parameters in the 2PL UIRT model using MML, Gibbs sampling, and NUTS when *N*=100.

	True		MML		Gibbs sampling		NUTS	
Item	*a*	*b*	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
1	1.461	-1.806	1.206	-1.880	1.369	-1.678	1.381	-1.664
2	0.928	-1.507	2.429	-0.963	1.704	-1.058	1.699	-1.056
3	0.346	0.421	-0.015	-13.127	0.193	0.403	0.193	0.417
4	0.017	1.432	-0.095	-0.846	0.136	0.161	0.134	0.161
5	1.085	-1.785	0.431	-3.994	0.900	-1.960	0.905	-1.955
6	1.823	-1.368	1.200	-1.808	1.340	-1.621	1.340	-1.624
7	0.685	0.328	0.657	-0.542	0.594	-0.490	0.588	-0.497
8	1.486	-0.910	1.558	-0.743	1.335	-0.768	1.328	-0.764
9	0.785	-1.133	0.645	-1.126	0.721	-0.949	0.721	-0.948
10	1.428	0.502	2.925	0.647	1.768	0.859	1.751	0.864
MSE			0.554	19.492	0.107	0.278	0.105	0.280

Table A4. The true and estimated values of the item parameters in the 2PL UIRT model using MML, Gibbs sampling, and NUTS when *N*=1000 and extremely easy items.

	True		MML		Gibbs sampling		NUTS	
Item	*a*	*b*	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
1	0.995	-1.916	1.086	-1.873	1.111	-1.835	1.109	-1.836
2	0.263	-1.901	0.256	-2.078	0.317	-1.669	0.320	-1.658
3	1.714	-1.937	1.948	-1.657	1.898	-1.660	1.908	-1.655
4	1.543	-1.980	1.118	-2.367	1.176	-2.277	1.176	-2.276
5	1.362	-1.976	1.400	-1.921	1.426	-1.887	1.423	-1.889
6	1.445	-1.970	1.862	-1.846	1.845	-1.837	1.837	-1.840
7	1.514	-1.910	1.754	-1.788	1.737	-1.781	1.742	-1.775
8	1.212	-1.956	0.961	-2.251	1.014	-2.154	1.016	-2.152
9	0.909	-1.958	0.952	-1.896	0.979	-1.850	0.982	-1.846
10	1.538	-1.995	1.641	-1.984	1.644	-1.967	1.644	-1.965
MSE			0.055	0.039	0.045	0.032	0.045	0.033

Table A5. The true and estimated values of the item parameters in the 2PL UIRT model using MML, Gibbs sampling, and NUTS when *N*=1000 and extremely difficult items.

	True		MML		Gibbs sampling		NUTS	
Item	*a*	*b*	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
1	0.710	1.972	0.908	1.740	0.927	1.706	0.930	1.703
2	1.910	1.949	2.378	1.792	2.224	1.821	2.238	1.815
3	1.380	1.986	1.571	1.894	1.587	1.872	1.589	1.869
4	1.217	1.960	1.439	1.834	1.454	1.810	1.447	1.816
5	0.181	1.930	0.238	1.132	0.258	1.038	0.261	1.023
6	1.563	1.942	1.806	1.832	1.787	1.826	1.790	1.824
7	0.440	1.938	0.270	3.102	0.385	2.189	0.384	2.198
8	1.732	1.925	1.389	2.180	1.408	2.155	1.418	2.142
9	0.781	1.972	0.895	1.791	0.917	1.755	0.927	1.738
10	1.706	1.948	1.784	1.941	1.757	1.943	1.762	1.939
MSE			0.057	0.220	0.043	0.109	0.044	0.113

APPENDIX B

Table B1. ANOVA effect sizes ($\hat{\omega}^2$) for log*RMSE* in estimating the discrimination (*a*) parameters in the 2PL UIRT model.

Variable	*df*	Sum of Squares	$\hat{\omega}^2$
Prior specifications for a_j (*P*)	2	1.364	0.015
Test length (*K*)	2	2.817	0.031
Sample size (*N*)	3	40.950	0.462
Algorithm (*A*)	1	0.249	0.003
P×*K*	4	0.099	0.000
P×*N*	6	1.549	0.016
P×*A*	2	0.491	0.005
K×*N*	6	0.159	0.001
K×*A*	2	0.061	0.000
N×*A*	3	0.615	0.006
P×*K*×*N*	12	0.337	0.002
P×*K*×*A*	4	0.121	0.001
P×*N*×*A*	6	1.228	0.013
K×*N*×*A*	6	0.145	0.001
P×*K*×*N*×*A*	12	0.284	0.001
Residual	1608	23.838	
Total	1679	88.509	

Table B2. ANOVA effect sizes ($\hat{\omega}^2$) for log*MAE* in estimating the discrimination (*a*) parameters in the 2PL UIRT model.

Variable	*df*	Sum of Squares	$\hat{\omega}^2$
Prior specifications for a_j (*P*)	2	1.157	0.015
Test length (*K*)	2	2.190	0.029
Sample size (*N*)	3	37.055	0.494
Algorithm (*A*)	1	0.034	0.000
P×*K*	4	0.090	0.001
P×*N*	6	0.876	0.011
P×*A*	2	0.066	0.001
K×*N*	6	0.129	0.001
K×*A*	2	0.025	0.000
N×*A*	3	0.269	0.003
P×*K*×*N*	12	0.152	0.000
P×*K*×*A*	4	0.049	0.000
P×*N*×*A*	6	0.540	0.006
K×*N*×*A*	6	0.076	0.000
P×*K*×*N*×*A*	12	0.152	0.000
Residual	1608	18.918	
Total	1679	74.904	

Table B3. ANOVA effect sizes ($\widehat{\omega}^2$) for log*RMSE* in estimating the difficulty (b) parameters in the 2PL UIRT model.

Variable	*df*	Sum of Squares	$\widehat{\omega}^2$
Prior specifications for a_j (*P*)	2	0.208	0.002
Test length (*K*)	2	0.558	0.006
Sample size (*N*)	3	18.811	0.228
Algorithm (*A*)	1	0.016	0.000
P×*K*	4	0.202	0.001
P×*N*	6	0.281	0.001
P×*A*	2	0.029	0.000
K×*N*	6	0.440	0.003
K×*A*	2	0.005	0.000
N×*A*	3	0.007	0.000
P×*K*×*N*	12	1.213	0.010
P×*K*×*A*	4	0.007	0.000
P×*N*×*A*	6	0.012	0.000
K×*N*×*A*	6	0.006	0.000
P×*K*×*N*×*A*	12	0.009	0.000
Residual	1608	49.110	
Total	1679	81.909	

Table B4. ANOVA effect sizes ($\widehat{\omega}^2$) for log*MAE* in estimating the difficulty (b) parameters in the 2PL UIRT model.

Variable	*df*	Sum of Squares	$\widehat{\omega}^2$
Prior specifications for a_j (*P*)	2	0.224	0.003
Test length (*K*)	2	0.488	0.007
Sample size (*N*)	3	19.112	0.285
Algorithm (*A*)	1	0.007	0.000
P×*K*	4	0.158	0.001
P×*N*	6	0.121	0.000
P×*A*	2	0.012	0.000
K×*N*	6	0.402	0.004
K×*A*	2	0.004	0.000
N×*A*	3	0.007	0.000
P×*K*×*N*	12	0.812	0.008
P×*K*×*A*	4	0.005	0.000
P×*N*×*A*	6	0.011	0.000
K×*N*×*A*	6	0.005	0.000
P×*K*×*N*×*A*	12	0.008	0.000
Residual	1608	34.482	
Total	1679	66.733	

Table B5. ANOVA effect sizes ($\widehat{\omega}^2$) for correlations between θ and $\hat{\theta}$ in the 2PL UIRT model.

Variable	*df*	Sum of Squares	$\widehat{\omega}^2$
Prior specifications for a_j (*P*)	2	0.005	0.000
Test length (*K*)	2	4.153	0.640
Sample size (*N*)	3	0.013	0.001
Algorithm (*A*)	1	0.000	0.000
P×*K*	4	0.005	0.000
P×*N*	6	0.020	0.002
P×*A*	2	0.000	0.000
K×*N*	6	0.009	0.000
K×*A*	2	0.000	0.000
N×*A*	3	0.000	0.000
P×*K*×*N*	12	0.062	0.007
P×*K*×*A*	4	0.000	0.000
P×*N*×*A*	6	0.000	0.000
K×*N*×*A*	6	0.000	0.000
P×*K*×*N*×*A*	12	0.000	0.000
Residual	1188	1.350	
Total	1259	6.487	

Table B6. ANOVA effect sizes ($\widehat{\omega}^2$) for log*RMSE* in estimating the discrimination (a_1) parameters in the 2PL multi-unidimensional IRT model.

Variable	*df*	Sum of Squares	$\widehat{\omega}^2$
Test length (*K*)	2	2.808	0.117
Sample size (*N*)	3	5.784	0.241
Algorithm (*A*)	1	0.000	0.000
Intertrait correlation (ρ)	2	0.541	0.021
K×*N*	6	0.283	0.008
K×*A*	2	0.003	0.000
K×ρ	4	0.206	0.006
N×*A*	3	0.000	0.000
N×ρ	6	0.245	0.007
A×ρ	2	0.004	0.000
K×*N*×*A*	6	0.008	0.000
K×*N*×ρ	12	0.368	0.008
K×*A*×ρ	4	0.006	0.000
N×*A*×ρ	6	0.004	0.000
K×*N*×*A*×ρ	12	0.007	0.000
Residual	768	11.106	
Total	839	23.814	

Table B7. ANOVA effect sizes ($\widehat{\omega}^2$) for log*MAE* in estimating the discrimination (a_1) parameters in the 2PL multi-unidimensional IRT model.

Variable	*df*	Sum of Squares	$\widehat{\omega}^2$
Test length (*K*)	2	2.897	0.115
Sample size (*N*)	3	6.732	0.267
Algorithm (*A*)	1	0.000	0.000
Intertrait correlation (ρ)	2	0.647	0.025
K×N	6	0.260	0.007
K×A	2	0.003	0.000
K×ρ	4	0.233	0.007
N×A	3	0.000	0.000
N×ρ	6	0.246	0.006
A×ρ	2	0.003	0.000
K×N×A	6	0.008	0.000
K×N×ρ	12	0.369	0.008
K×A×ρ	4	0.005	0.000
N×A×ρ	6	0.004	0.000
K×N×A×ρ	12	0.006	0.000
Residual	768	10.935	
Total	839	25.032	

Table B8. ANOVA effect sizes ($\widehat{\omega}^2$) for log*RMSE* in estimating the discrimination (a_2) parameters in the 2PL multi-unidimensional IRT model.

Variable	*df*	Sum of Squares	$\widehat{\omega}^2$
Test length (*K*)	2	2.841	0.116
Sample size (*N*)	3	7.651	0.314
Algorithm (*A*)	1	0.001	0.000
Intertrait correlation (ρ)	2	0.159	0.006
K×N	6	0.171	0.004
K×A	2	0.001	0.000
K×ρ	4	0.129	0.003
N×A	3	0.001	0.000
N×ρ	6	0.331	0.011
A×ρ	2	0.001	0.000
K×N×A	6	0.005	0.000
K×N×ρ	12	0.394	0.010
K×A×ρ	4	0.004	0.000
N×A×ρ	6	0.001	0.000
K×N×A×ρ	12	0.004	0.000
Residual	768	9.191	
Total	839	24.248	

Table B9. ANOVA effect sizes ($\hat{\omega}^2$) for log*MAE* in estimating the discrimination (a_2) parameters in the 2PL multi-unidimensional IRT model.

Variable	*df*	Sum of Squares	$\hat{\omega}^2$
Test length (*K*)	2	2.900	0.111
Sample size (*N*)	3	8.062	0.309
Algorithm (*A*)	1	0.000	0.000
Intertrait correlation (ρ)	2	0.112	0.003
$K\times N$	6	0.210	0.005
$K\times A$	2	0.001	0.000
$K\times\rho$	4	0.133	0.003
$N\times A$	3	0.001	0.000
$N\times\rho$	6	0.301	0.009
$A\times\rho$	2	0.001	0.000
$K\times N\times A$	6	0.006	0.000
$K\times N\times\rho$	12	0.350	0.007
$K\times A\times\rho$	4	0.005	0.000
$N\times A\times\rho$	6	0.001	0.000
$K\times N\times A\times\rho$	12	0.007	0.000
Residual	768	10.260	
Total	839	25.978	

Table B10. ANOVA effect sizes ($\hat{\omega}^2$) for log*RMSE* in estimating the difficulty (b_1) parameters in the 2PL multi-unidimensional IRT model.

Variable	*df*	Sum of Squares	$\hat{\omega}^2$
Test length (*K*)	2	0.562	0.014
Sample size (*N*)	3	4.714	0.129
Algorithm (*A*)	1	0.000	0.000
Intertrait correlation (ρ)	2	0.013	0.000
$K\times N$	6	0.614	0.012
$K\times A$	2	0.001	0.000
$K\times\rho$	4	0.275	0.004
$N\times A$	3	0.000	0.000
$N\times\rho$	6	0.375	0.005
$A\times\rho$	2	0.002	0.000
$K\times N\times A$	6	0.002	0.000
$K\times N\times\rho$	12	0.843	0.013
$K\times A\times\rho$	4	0.002	0.000
$N\times A\times\rho$	6	0.003	0.000
$K\times N\times A\times\rho$	12	0.003	0.000
Residual	768	24.889	
Total	839	35.664	

Table B11. ANOVA effect sizes ($\widehat{\omega}^2$) for log*MAE* in estimating the difficulty (b_1) parameters in the 2PL multi-unidimensional IRT model.

Variable	*df*	Sum of Squares	$\widehat{\omega}^2$
Test length (*K*)	2	0.747	0.023
Sample size (*N*)	3	5.750	0.185
Algorithm (*A*)	1	0.000	0.000
Intertrait correlation (ρ)	2	0.000	0.000
K×*N*	6	0.603	0.015
K×*A*	2	0.001	0.000
K×ρ	4	0.213	0.004
N×*A*	3	0.000	0.000
N×ρ	6	0.459	0.010
A×ρ	2	0.001	0.000
K×*N*×*A*	6	0.004	0.000
K×*N*×ρ	12	0.539	0.008
K×*A*×ρ	4	0.002	0.000
N×*A*×ρ	6	0.002	0.000
K×*N*×*A*×ρ	12	0.003	0.000
Residual	768	18.273	
Total	839	30.710	

Table B12. ANOVA effect sizes ($\widehat{\omega}^2$) for log*RMSE* in estimating the difficulty (b_2) parameters in the 2PL multi-unidimensional IRT model.

Variable	*df*	Sum of Squares	$\widehat{\omega}^2$
Test length (*K*)	2	1.054	0.023
Sample size (*N*)	3	7.910	0.186
Algorithm (*A*)	1	0.000	0.000
Intertrait correlation (ρ)	2	0.047	0.000
K×*N*	6	0.202	0.000
K×*A*	2	0.000	0.000
K×ρ	4	0.534	0.009
N×*A*	3	0.000	0.000
N×ρ	6	0.255	0.001
A×ρ	2	0.000	0.000
K×*N*×*A*	6	0.001	0.000
K×*N*×ρ	12	0.669	0.006
K×*A*×ρ	4	0.001	0.000
N×*A*×ρ	6	0.003	0.000
K×*N*×*A*×ρ	12	0.004	0.000
Residual	768	26.736	
Total	839	41.911	

Table B13. ANOVA effect sizes ($\widehat{\omega}^2$) for log*MAE* in estimating the difficulty (b_2) parameters in the 2PL multi-unidimensional IRT model.

Variable	*df*	Sum of Squares	$\widehat{\omega}^2$
Test length (*K*)	2	1.327	0.034
Sample size (*N*)	3	8.588	0.226
Algorithm (*A*)	1	0.000	0.000
Intertrait correlation (ρ)	2	0.015	0.000
K×N	6	0.309	0.004
K×A	2	0.000	0.000
K×ρ	4	0.415	0.008
N×A	3	0.000	0.000
N×ρ	6	0.219	0.001
A×ρ	2	0.000	0.000
K×N×A	6	0.002	0.000
K×N×ρ	12	0.703	0.010
K×A×ρ	4	0.001	0.000
N×A×ρ	6	0.003	0.000
K×N×A×ρ	12	0.004	0.000
Residual	768	20.799	
Total	839	37.651	

Table B14. ANOVA effect sizes ($\widehat{\omega}^2$) for correlations between θ_1 and $\hat{\theta}_1$ in the 2PL multi-unidimensional IRT model.

Variable	*df*	Sum of Squares	$\widehat{\omega}^2$
Test length (*K*)	2	4.748	0.670
Sample size (*N*)	3	0.007	0.000
Algorithm (*A*)	1	0.000	0.000
Intertrait correlation (ρ)	2	0.449	0.063
K×N	6	0.009	0.000
K×A	2	0.000	0.000
K×ρ	4	0.147	0.019
N×A	3	0.000	0.000
N×ρ	6	0.004	0.000
A×ρ	2	0.000	0.000
K×N×A	6	0.000	0.000
K×N×ρ	12	0.045	0.002
K×A×ρ	4	0.000	0.000
N×A×ρ	6	0.000	0.000
K×N×A×ρ	12	0.000	0.000
Residual	648	1.665	
Total	719	7.074	

Table B15. ANOVA effect sizes ($\widehat{\omega}^2$) for correlations between θ_2 and $\hat{\theta}_2$ in the 2PL multi-unidimensional IRT model.

Variable	*df*	Sum of Squares	$\widehat{\omega}^2$
Test length (*K*)	2	3.831	0.616
Sample size (*N*)	3	0.021	0.002
Algorithm (*A*)	1	0.000	0.000
Intertrait correlation (ρ)	2	0.416	0.066
K×*N*	6	0.042	0.004
K×*A*	2	0.000	0.000
K×ρ	4	0.079	0.011
N×*A*	3	0.000	0.000
N×ρ	6	0.040	0.004
A×ρ	2	0.000	0.000
K×*N*×*A*	6	0.000	0.000
K×*N*×ρ	12	0.035	0.000
K×*A*×ρ	4	0.000	0.000
N×*A*×ρ	6	0.000	0.000
K×*N*×*A*×ρ	12	0.000	0.000
Residual	648	1.752	
Total	719	6.216	

VITA

Graduate School
Southern Illinois University

Meng-I Chang

chang.mengi@gmail.com

Chung Yuan Christian University, Taiwan
Bachelor of Science, Electronic Engineering, June, 2001

Southern Illinois University Carbondale
Master of Arts, Teaching English to Speakers of Other Languages, May, 2008

Special Honors and Awards:
Patricia Borgsmiller Elmore and Donald E. Elmore Doctoral Scholar Award 2015, 2016
SIUC graduate tuition award 2007

Dissertation Title:
A Comparison of Two MCMC Algorithms for Estimating the 2PL IRT Models

Major Professor: Yanyan Sheng

Publications:
Chang, M. I., & Sheng, Y. (in press). A comparison of two MCMC algorithms for the 2PL IRT model. In R. E. Millsap, L. A. van der Ark, S. A. Culpepper, J. A. Douglas, W. C. Wang, & M. Wiberg (Eds.), New Developments in Quantitative Psychology. New York: Springer.