1-1-2019

A Comparison of Bayesian Estimation Techniques in a Multidimensional Two-Parameter Partial Credit Item Response Model

Peiyan Liu
University of Denver

A Comparison of Bayesian Estimation Techniques in a Multidimensional Two-Parameter Partial Credit Item Response Model

Abstract

Bayesian estimation methods have shown better performance than the traditional Marginal Maximum Likelihood (MML) estimation method for parameter estimation in relatively simple item response models. However, extant literature is lacking on the investigation of Bayesian parameter estimation approaches for a multidimensional two parameter partial credit (M2PPC) model, therefore this simulation study investigated the performance of two Bayesian Markov Chain Monte Carlo (MCMC) algorithms: Gibbs Sampler and Hamiltonian Monte Carlo-No-U-Turn-Sampler (HMC-NUTS) for M2PPC models' parameter estimation. It compared the estimation accuracy and computing speed in different combinations of situations, including prior choices, test lengths, and the relationships between dimensions.

The datasets were generated based on the distributions from existing literature, and the conditions were fully crossed. It ended up with 36 conditions: Bayesian MCMC algorithms (Gibbs sampler and HMC-NUTS), prior choices (Matched Prior, Vague Prior and Hierarchical Prior), test lengths (15 and 30), and the relationships between dimensions (low = .2, medium = .5, and high = .8). Root Mean Squared Errors (RMSE) and Bias for each of the recovered parameter in all the conditions were calculated. Sets of four-way ANOVAs were conducted to check the contribution of the four factors--Bayesian algorithm, prior choice, test length, and interdimensional correlation--to the total variance in RMSE and Bias. The computational speed was also recorded for each of the estimations.

The first finding is that when considering the computational speed and estimation accuracy, the results of parameter recovery of the M2PPC model show that Gibbs Sampler and HMC-NUTS performed similarly in all the simulated conditions. The second finding is concerning test length. The precision of item parameter estimates increased as the test length decreased, but the accuracy of person parameter estimates increased as the test length increased in all the simulated conditions for both Gibbs Sampler and HMC-NUTS. Test length had no consistent impact on Bias for either item parameter or person parameter estimates. The third finding is that different interdimensional correlations did not influence the recovery of item parameters but affected the precision of the estimation of person parameters. The accuracy of the person parameter recovery increased as the interdimensional correlation increased in all the different conditions for both Gibbs Sampler and HMC-NUTS. The results of analyses of variance (ANOVAs) supported the previous conclusions. This dissertation study concluded with limitations and recommendations for future work.

Document Type

Dissertation

Degree Name

Ph.D.

Department

Quantitative Research Methods

First Advisor

Kathy Green, Ph.D.

Second Advisor

Cathrine Durso

A COMPARISON OF BAYESIAN ESTIMATION TECHNIQUES IN A MULTIDIMENSIONAL TWO-PARAMETER PARTIAL CREDIT ITEM RESPONSE MODEL

A Dissertation

Presented to

the Faculty of the Morgridge College of Education

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Peiyan Liu

March 2019

Advisor: Kathy Green, Ph.D

Author: Peiyan Liu
Title: A COMPARISON OF BAYESIAN ESTIMATION TECHNIQUES IN A MULTIDIMENSIONAL TWO-PARAMETER PARTIAL CREDIT ITEM RESPONSE MODEL
Advisor: Kathy Green, Ph.D
Degree Date: March 2019

Abstract

Bayesian estimation methods have shown better performance than the traditional Marginal Maximum Likelihood (MML) estimation method for parameter estimation in relatively simple item response models. However, extant literature is lacking on the investigation of Bayesian parameter estimation approaches for a multidimensional two parameter partial credit **(**M2PPC) model, therefore this simulation study investigated the performance of two Bayesian Markov Chain Monte Carlo (MCMC) algorithms: Gibbs Sampler and Hamiltonian Monte Carlo-No-U-Turn-Sampler (HMC-NUTS) for M2PPC models' parameter estimation. It compared the estimation accuracy and computing speed in different combinations of situations, including prior choices, test lengths, and the relationships between dimensions.

The datasets were generated based on the distributions from existing literature, and the conditions were fully crossed. It ended up with 36 conditions: Bayesian MCMC algorithms (Gibbs sampler and HMC-NUTS), prior choices (Matched Prior, Vague Prior and Hierarchical Prior), test lengths (15 and 30), and the relationships between dimensions (low = .2, medium = .5, and high = .8). Root Mean Squared Errors (RMSE) and Bias for each of the recovered parameter in all the conditions were calculated. Sets of four-way ANOVAs were conducted to check the contribution of the four factors--Bayesian algorithm, prior choice, test length, and interdimensional correlation--to the

total variance in RMSE and Bias. The computational speed was also recorded for each of the estimations.

The first finding is that when considering the computational speed and estimation accuracy, the results of parameter recovery of the M2PPC model show that Gibbs Sampler and HMC-NUTS performed similarly in all the simulated conditions. The second finding is concerning test length. The precision of item parameter estimates increased as the test length decreased, but the accuracy of person parameter estimates increased as the test length increased in all the simulated conditions for both Gibbs Sampler and HMC-NUTS. Test length had no consistent impact on Bias for either item parameter or person parameter estimates. The third finding is that different interdimensional correlations did not influence the recovery of item parameters but affected the precision of the estimation of person parameters. The accuracy of the person parameter recovery increased as the interdimensional correlation increased in all the different conditions for both Gibbs Sampler and HMC-NUTS. The results of analyses of variance (ANOVAs) supported the previous conclusions. This dissertation study concluded with limitations and recommendations for future work.
Keywords: Bayesian estimation, Gibbs sampler, HMC-NUTS, RMSE, Bias

Acknowledgements

First of all I give sincere thanks to my advisor, Prof. Kathy Green, who has not only been a mentor but also a friend who cares and believes in me. Her guidance and gifted personality have made this rewarding journey possible in many ways. She is a remarkable role model who inspires my academic, career, and personal lives. I would like to extend my gratitude to one of my committee members, Prof. Cathrine Durso, whose advice and insight on Bayesian statistics guide me through the hard scripting time of my dissertation. I also wanted to thank my another committee member Prof. Duan Zhang for her helpful insights and unlimited supports during my whole doctoral endeavor. I owe many thanks to the faculty, to my fellow students, and to the kind and supportive staff of the University of Denver. I was treated with all the love and respect. I would like to express my thanks to the professors and students who took part in one way or another in my various studies, and generously shared your time, ideas and passion with me. My appreciation would also go to my dear friends Yanqiu Zhang and Dr. Lin Ma who are my greatest advocates and willing to listen to my sadness and happiness all the dissertation way.

Last but not least, very special thanks go to my parents even though they are 7,000 miles away. I am so grateful that I have a great mom who always encourages me to chase my dream with passion, no matter what kind of setbacks I encounter. Her unconditional love shapes who I am. My incredible dad with his humble and hardworking traits motivate my perseverance to stick to what I have started to the very end and ultimately to make my dream come true.

Table of Contents

Chapter One: Introduction and Literature Review……..…....1
Introduction……..1
Problem Statement……..…..2
Literature Review……..…3
Item Response Theory (IRT) Models……..…3
Frequentist Inference via Marginal Maximum Likelihood (MML)……7
Bayesian Markov Chain Monte Carlo (MCMC)……...…..9
Gibbs Sampler……...…….11
Hamiltonian Monte Carlo-No-U-Turn-Sampler (HMC-NUTS)...13
Summary of Comparison Studies of Estimation Methods in IRT…….16
Purpose of the Study……..18
Delimitations……..19
Definition of Terms……..…..20

Chapter Two: Method
Introduction……23
Monte Carlo Simulation…….……...…….23
Analysis Procedures……..…….24
Phase I Data Generation…….24
Phase II Parameter Estimation with Different Priors…….....26
Phase III Outcome Analysis: Parameter Recovery Evaluation Criteria....29

Chapter Three: Results……...……38
Introduction……...…….38
Convergence of Markov Chains……...….39
Item Parameter Recovery……..…..……45
Person Parameter Recovery…….…….…64
Computational Speed of Gibbs Sampler and HMC-NUTS……...73
Analysis of Variance (ANOVA) Results…….…..74

Chapter Four: Discussion and Conclusion……...…….……82
Introduction…….…..82
The Findings of This Study……...……82
Discussions……..…..…….85
Limitations……87
Directions for Future Studies……..……..…….88

References……...…….…91

Appendix A……...……96
R*jags* Script for Estimation in the M2PPC Model……..……..96
Appendix B……...……...……..……101

Rstan Script for Estimation in the M2PPC Model……..….....101
Appendix C……...……..…….....107
Rationale for NOT Needing IRB Review……....107

List of Tables

Chapter Two
Table 1 Simulation Design…….….32

Chapter Three
Table 2 RMSE and Bias of Intercept Estimates for the Four Transitional Points: γ(1), γ(2), γ(3), and γ(4) in the M2PPC Model When Test Length = 30..56
Table 3 RMSE and Bias of Slope Estimates on the Three Dimensions: a(1), a(2), and a(3) in the M2PPC Model When Test Length = 30……....58
Table 4 RMSE and Bias of Intercept Estimates for the Four Transitional Points: γ(1), γ(2), γ(3), and γ(4) in the M2PPC Model When Test Length = 15...60
Table 5 RMSE and Bias of Slope Estimates on the Three Dimensions: a(1), a(2), and a(3) in the M2PPC Model When Test Length = 15……....62
Table 6 RMSE and Bias of Person Ability Estimates on Three Dimensions: θ(1), θ(2), and θ(3) in the M2PPC Model When Test Length = 30……..…....69
Table 7 RMSE and Bias of Person Ability Estimates on Three Dimensions: θ(1), θ(2), and θ(3) in the M2PPC Model When Test Length = 15……...71
Table 8 Effect Sizes (ω^2) for Main Effects and Interactions on LogRMSE of Item Parameter Estimates in the M2PPC Model…….…...75
Table 9 Effect Sizes (ω^2) for Main Effects and Interactions on LogRMSE of Person Parameter Estimates in the M2PPC Model……...……76
Table 10 Effect Sizes (ω^2) for Main Effects and Interactions on LogBias of Item Parameter Estimates in the M2PPC Model……78
Table 11 Effect Sizes (ω^2) for Main Effects and Interactions on LogBias of Person Parameter Estimates in the M2PPC Model……79

List of Figures

Chapter Two

Figure 1 Visual Representation of Analysis Procedure ……..…….37

Chapter Three

Figure 2 Trace Plots of the Intercept ($\gamma(1)$, $\gamma(2)$, $\gamma(3)$ and $\gamma(4)$), Slope Parameter ($a(1)$, $a(2)$, and $a(3)$), and Person Ability Parameter ($\theta(1)$, $\theta(2)$, and $\theta(3)$) in the M2PPC Model Using Gibbs Sampler (upper panel) and HMC-NUTS (lower panel)……..…….…….….. 42

Figure 3 Gelman-Rubin of the Intercept ($\gamma(1)$, $\gamma(2)$, $\gamma(3)$ and $\gamma(4)$), Slope Parameter ($a(1)$, $a(2)$, and $a(3)$), and Person Ability Parameter ($\theta(1)$, $\theta(2)$, and $\theta(3)$) in the M2PPC Model Using Gibbs Sampler (upper panel) and HMC-NUTS (lower panel)……..……..……44

Figure 4 Average RMSE for Recovering Intercept Parameter $\gamma(1)$ Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC model…….…….48

Figure 5 Average RMSE for Recovering Intercept Parameter $\gamma(2)$ Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC model…….…….49

Figure 6 Average RMSE for Recovering Intercept Parameter $\gamma(3)$ Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC model…….…….50

Figure 7 Average RMSE for Recovering Intercept Parameter $\gamma(4)$ Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC model…….……..……52

Figure 8 Average RMSE for Recovering Slope Parameter $a(1)$ Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC model…….…….....53

Figure 9 Average RMSE for Recovering Slope Parameter $a(2)$ Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC model…….…….55

Figure 10 Average RMSE for Recovering Slope Parameter $a(3)$ Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC model…….……..…..57

Figure 11 Average RMSE for Recovering Person Ability Parameter $\theta(1)$ Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and

15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC model…….67
Figure 12 Average RMSE for recovering person ability parameter $\theta(2)$ using Vague, Matched and Hierarchical Prior when Test Length equal to 30 and 15 and interdimensional correlation equal to .2, .5 and .8 in the M2PPC model…….68
Figure 13 Average RMSE for recovering person ability parameter $\theta(3)$ using Vague, Matched and Hierarchical Prior when Test Length equal to 30 and 15 and interdimensional correlation equal to .2, .5 and .8 in the M2PPC model……..69

Chapter One：Introduction and Literature Review

Chapter One provides an introduction to this study and a brief overview of the problem. After the introduction to the topic, the literature review begins with an overview of IRT models, and introduces multidimensional two parameter partial credit (M2PPC) models. I then present the most commonly used frequentist parameter estimation method, marginal maximum likelihood (MML), in IRT models. Subsequently, Bayesian inference for item response models' parameter estimation is described. In this study, I focus primarily on Bayesian Markov Chain Monte Carlo (MCMC) techniques. The literature introduces two Bayesian estimation techniques--Gibbs Sampler and Hamiltonian Monte Carlo-No-U-Turn-Sampler (HMC-NUTS)--for IRT models. A summary of comparative studies of how Bayesian MCMC estimation techniques behave in IRT models concludes the literature review section.

Then, the rationale for this study is stated. Finally, I present the research purposes and research questions that I address through this study. I include several limitations and definitions of the terms of my study.

Introduction

M2PPC models apply to the investigation of the latent factors underlying psychological, educational, and medical tests and questionnaires composed of items with few response categories. M2PPC models can utilize all the information from each

response to better measure people to create psychometrically sound instruments. In general, IRT models have the features of establishing a probabilistic relationship between responses on a set of items and the latent traits/factors on the basis of a set of parameters (Tsutakawa & Lin, 1986). Therefore, the first and foremost issue associated with IRT models is appropriate parameter estimation, which lays the foundation for the further application of IRT models in different content areas.

This simulation study investigates the performance of two Bayesian MCMC algorithms: Gibbs Sampler and HMC-NUTS for M2PPC models' parameter estimation. It compares the estimation accuracy and computing speed in different combinations of situations, including prior distributions, test lengths, and the relationships between dimensions.

Problem Statement

The increase in computing power of computer hardware and the development of advanced psychometric software makes it possible to calibrate complex IRT models, such as M2PPC models. However, until recently researchers only compared the parameter estimation approaches for the multidimensional 2-PL dichotomous model (Martin-Fernandez, M., & Revuelta, J., 2017), unidimensional 1-PL and 2-PL models (Natesan et al., 2016), and multi-unidimensional (two dimensions) 2-PL graded response models (Kuo & Sheng, 2016), and there are concerns and confusions in the limited literature about under which conditions or under which combination of conditions (including test lengths, interdimensional correlations, and prior choices), which parameter estimation method(s) is more appropriate. In addition, no study was identified that compared the two

Bayesian MCMC estimation methods-Gibbs Sampler and HMC-NUTS estimation for an M2PPC model, but the two MCMC estimation methods have shown better performance than MML for the relatively simple IRT models examined in existing psychometric research (Martin-Fernandez & Revuelta, 2017; Natesan et al., 2016). Finally, the lack of identified literature on the investigation of Bayesian parameter estimation approaches for M2PPC models serves to motivate an exploration of how the two Bayesian MCMC algorithms behave in M2PPC models. The present study addressed this gap in the literature.

Literature Review

IRT Models

IRT begins with the proposition that a person's response to a certain item is determined by an unobservable attribute of that person. That attribute is referred to as an "ability" or "trait." Because abilities are not directly observed, they are called "latent traits" or "latent abilities." IRT models the relationship between persons' performances on a test item and their levels of performance on an overall measure of the ability that item was designed to reflect. Several different statistical models with different parameters are used to represent both item and person characteristics. Some of the models are used to quantify the probability of correct answers as a function of unobserved person abilities and other parameters to explain the difficulty and the discriminatory power of the items in the test. Some also include a threshold parameter for the probability of the correct answer to account for the guessing effect in multiple choice items. Suppose that each of N persons is given n items (questions). The response y_{ij} for the jth person and the ith item

is recorded as 1 or 0 according to whether the person answers the item correctly or incorrectly. Let p_{ij} denote the probability that the *j*th person can answer the *i*th item correctly. The model is the following 3-parameter logistic (PL) response function:

$$p_{ij}=P\,(y_{ij}=1|\,\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})= c_i+(1- c_i)\frac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}} \quad (1)$$

for i=1, 2, …, n, and j=1, 2, …, N, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$, with $-\infty < \theta_j < +\infty$, is a vector of ability variables for N persons; $\boldsymbol{a} = (a_1, \ldots, a_n)$, with $a_i > 0$, is a vector of the item discrimination parameters for n items; $\boldsymbol{b} = (b_1, \ldots, b_n)$, with $-\infty < b_i < +\infty$, is a vector of the item difficulty parameters for n items, and $\boldsymbol{c} = (c_1, \ldots, c_n)$, with $0 < c_i < 1$, is a vector of the item guessing parameters for n items. In particular, when c_i=0, model (1) reduces to a 2PL model.

There are three assumptions commonly made in IRT models. The first assumption is about the dimensional structure of test data. The models that assume a single ability are referred to as unidimensional models, such as model (1). However, in many situations, there is an *a priori* assumption that multiple abilities are involved in producing the responses (e.g., a math word problem may measure a composite skill of math and reading comprehension with varying emphases on the two skills). Such cases require multidimensional models, which is the focus of this study. The second assumption is local independence. This assumption states that a person's response to one item does not affect his/her response to any other items in the test: only the person's ability and the characteristics of the item can influence the response to that item. The last assumption is about the mathematical form of the item characteristic curve (ICC). The relationship

between a person's latent ability and the probability of the examinee correctly responding to a particular item is modeled by a mathematical function called the item characteristic function. It is the linear or nonlinear function for the regression of item score on the ability measured by the test (Hambleton, 1989).

As noted in the previous paragraph, it is often more realistic to hypothesize that a person's response to an item is due to his/her locations on multiple latent variables. Thus we have a multidimensional latent space. For instance, students' learning self-efficacy involves cognitive and affective dimensions, and the responses to items on a measure of self-efficacy are a function of the person's locations on these two dimensions. This example is different from the math word problem example, in which the person's highly developed reading comprehension ability can compensate for the lower math proficiency. In contrast, the person's location on the cognitive dimension of self-efficacy cannot compensate for the person's location on the affective dimension, and such situations present noncompensatory multidimensional models. (Details can be found in Whitely, 1980.) The math word problem example presents a case of a compensatory multidimensional model, which is the focus of this study.

Similarly to model (1), the multidimensional 3PL model is defined as:

$$p_{ij}=P\,(y_{ij}=1|\,\boldsymbol{\theta_j},\boldsymbol{a_i},\boldsymbol{b_i},\boldsymbol{c})=c_i+(1-c_i)\frac{\exp\{\sum_{k=1}^{m}[a_{ik}(\theta_{jk}-b_{ik})]\}}{\exp\{\sum_{k=1}^{m}[a_{ik}(\theta_{jk}-b_{ik})]\}+1} \qquad (2)$$

for i=1, 2, …, n, and j=1, 2, …, N, where $\boldsymbol{\theta_j}=(\theta_{j1},\dots,\theta_{jk},\dots,\theta_{jm})$ is a vector of m ability values, with $-\infty<\theta_{jk}<+\infty$, for j=1,2,…, N, k=1,2,…, m; $\boldsymbol{a_i}=(a_{i1},\dots,a_{ik},\dots,a_{im})$ is the vector of loadings of item i on the m abilities (item discriminations) with $a_{ik}>0$ for

i=1,2,…, n, k=1,2,…, m ; $\boldsymbol{b_i} = (b_{i1}, \ldots, b_{ik}, \ldots, b_{im})$ is the item difficulty parameter vector along the m abilities dimensions with $-\infty < b_{ik} < +\infty$ for i=1,2,…, n, k=1,2,…, m; and $\boldsymbol{c} = (c_1, \ldots, c_n)^T$ is the vector of the item guessing parameters for n items with $0 < c_i < 1$ *for* i=1,2,…, n.

Note that when m=1, model (2) reduces to a unidimensional 3PL model (1). When c_i=0, model (2) reduces to a multidimensional 2PL model (1).

$$p_{ij}=P\,(y_{ij}=1|\,\boldsymbol{\theta_j}, \boldsymbol{a_i}, \boldsymbol{b_i})=\frac{\exp\{\sum_{k=1}^{m}[a_{ik}(\theta_{jk}-b_{ik})]\}}{\exp\{\sum_{k=1}^{m}[a_{ik}(\theta_{jk}-b_{ik})]\}+1} \quad (3)$$

Letting $\gamma_i=-\sum_{k=1}^{m} a_{ik}b_{ik}$, equation (3) becomes

$$p_{ij}=P\,(y_{ij}=1|\,\boldsymbol{\theta_j}, \boldsymbol{a_i}, \boldsymbol{\gamma_i})=\frac{\exp(\sum_{k=1}^{m} a_{ik}\theta_{jk}+\gamma_i)}{\exp(\sum_{k=1}^{m} a_{ik}\theta_{jk}+\gamma_i)+1} \quad (4)$$

In practice, response types may not be limited to binary responses as in the previous models. For instance, rating scales where a person chooses a response from a set of choices are used to measure numerous educational, psychological, and medical outcomes. A "rating scale" model is one in which all items (or groups of items) share the same rating scale structure, such as Likert scales. For another example, partial scales, a partial credit model is one in which each item has a unique rating scale structure. Several models can score such data type, such as the graded response model, nominal response model, and generalized partial credit model, among which the M2PPC model (Muraki, 1997) is one of the most often used polytomous IRT models and is the focus of this study:

$P\ (y=x_{iq}|\ \boldsymbol{\theta}, \boldsymbol{a_i}, \boldsymbol{\gamma_i})= \frac{\exp[\sum_{h=0}^{q}(\sum_{k=1}^{m} a_{ik}\theta_k+\gamma_{ih})]}{\sum_{t=0}^{p}\{\exp[\sum_{h=0}^{t}(\sum_{k=1}^{m} a_{ik}\theta_k+\gamma_{ih})]\}}$, setting $a_{ik}\theta_k + \gamma_{i0}$ =0 as a notational convenience. (5)

Suppose that model (5) describes the probability that one person's response is in item *i*'s *q* category, x_{iq}. The vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k, \ldots, \theta_m)$ is the vector of the *m* ability parameters for that person, with $-\infty < \theta_k < +\infty$ for *k*=1,2,…, *m*; $\boldsymbol{a_i} = (a_{i1}, \ldots, a_{ik}, \ldots, a_{im})$ the vector of loadings (item discriminations) of item *i* on the *m* abilities, with $a_{ik} > 0$ for *i*=1, 2, …, *n*, *k*=1,2,…, *m* ; $\boldsymbol{\gamma_i} = (\gamma_{i0}, \ldots, \gamma_{iq}, \ldots, \gamma_{ip})$ a vector of intercept parameters reflecting the interaction of the transition location parameters and discrimination parameters ($\gamma_{ih}=-\sum_{k=1}^{m} a_{ik}b_{ikh}$), with $-\infty < \gamma_{iq} < +\infty$ for *i*=1, 2, …, *n* and *q*=0, 1,…, *p*. In a proficiency measurement situation, γ_{ih} can be interpreted as related to an item's difficulty though with an opposite sign. Following Muraki (1997), this defines that first boundary location as zero, thus there are *p* transition locations and *p+1* categories.

Frequentist Inference via Marginal Maximum Likelihood (MML)

Accurate parameter estimation is a crucial problem in item response theory (IRT), and currently frequentist inference via MML, developed by Bock and Aitkin (1981), is the most widely used parameter estimation method in item response models (Martin-Fernandez & Revuelta, 2017). MML has a two-step procedure. After obtaining the joint probability of the item response vector given the person parameters, MML treats persons as random effects and derives a marginal probability of observing the item response vector by integrating the person effect out of the joint likelihood to separate item parameters from person parameters, thereby the IRT characteristic of "sample free."

Hence, in MML, item parameters can be obtained using an expectation-maximization (EM) algorithm, and person parameters can be iteratively estimated using the item parameters. Bock and Aitkin's MML is limited to IRT models with lower dimensions, since it uses fixed Gauss-Hermite quadrature (Baker & Kim, 2004). As the number of dimensions increases, the number of quadrature points increases exponentially, which need to be accounted for by decreasing the number of quadrature points in each dimension. Schilling and Bock (2005) proposed using adaptive quadrature for better accuracy when a relatively small number of quadrature points are used for each dimension. This estimation method can be used with a moderate number of dimensions (e.g., 3-4 dimensions) in item response models.

The MML estimator is a function of the data that has a distribution. This renders the estimator a random variable, and the realization of the random variable, the MML estimates, are obtained from a sample of data from the population. Standard errors of these estimates capture the uncertainty of these estimates, and are used to construct confidence intervals. The estimation routines heavily depend on asymptotic arguments to justify the calculation of parameter estimates, standard errors, and the sampling distribution of the parameter estimates. Moreover, these parameter values are treated as fixed or constant, which makes it inappropriate to examine their probabilities. The standard error is a measure of the variability of the value of the parameter estimator due to sampling data from the population. Likewise, the probabilistic interpretation of the confidence interval relies on the sampling distribution of the interval on repeated sampling of data, and applies to the process of interval estimator construction. These

notions refer to the variability of a parameter estimator, which is a distribution of parameter estimates in repeated sampling. In other words, the parameters are assumed to be constant over the repeated samples from the data. In MML, the probabilistic statements refer to the variability and likely values of the parameter estimator rather than the parameters themselves. I discuss MML here only to provide a background for Bayesian inference, and a more detailed description of MML can be found in Schilling and Bock (2005).

Bayesian Markov Chain Monte Carlo (MCMC)

Advances in computational statistics in recent decades have made Bayesian estimation plausible for IRT models parameter estimation. Bayesian inference shares some features with frequentist inference. Both approaches treat the data as random and assign the data a distribution, which is conditional on model parameters--$p(x|\theta)$. The likelihood of the possible data is computed given the possible parameter values. Once the data are observed, the observed data are taken as a function of the model parameters. The function is the likelihood function of the data. Thus, the likelihood function plays as vital a role in Bayesian inference as it does in MML estimation.

However, Bayesian inference is different from frequentist inference in the following way. The frequentist approach treats parameters as fixed, but the Bayesian approach treats parameters as random. They are random in the sense that people have uncertain knowledge about them. Different distributions model people's beliefs about them before and after analysis of relevant data. In a Bayesian approach, the model parameters θ are assigned *prior* distributions $p(\theta)$ which reflects the researcher's beliefs,

or prior knowledge, about the parameters. Bayesian inference then synthesizes the prior distribution and the likelihood of the data via Bayes' theorem to yield the posterior distribution of the parameters, $p(\theta|x)$:

$$p(\theta|x)=\frac{p(x,\theta)}{p(x)} \tag{6}$$

$$= \frac{p(x|\theta)p(\theta)}{p(x)}$$

$$\propto p(x|\theta)p(\theta)$$

Note: θs in this section refer to model parameters of interest not the person parameter in IRT models.

Bayesian estimates are asymptotically distribution-free and depend less on the distribution of the data, which is one advantage of Bayesian inference over traditional estimation techniques (Ansari & Jedidi, 2000). Besides, Bayesian methods have a potential advantage over MML in small samples and when the examinee ability distribution is not normal. In a nutshell, the purpose of Bayesian inference is to maintain a full posterior probability distribution over a set of random variables.

Bayesian simulation algorithms consist of drawing samples of parameters from the posterior distribution using MCMC algorithms, which avoid the complicated derivatives and are cost-efficient concerning computational time. After obtaining the posterior distribution, we can compute any statistics on it based on the simulated samples from that distribution. One MCMC algorithm with more extensive research in the Bayesian IRT literature is the Gibbs Sampler. The HMC-NUTS (Hoffman and Gelman, 2014) was introduced recently to overcome some of the shortcomings of the previous

algorithms, such as speeding up convergence of MCMC chains (Hoffman and Gelman, 2014), whose performance including model fit indices and computational speed has not been thoroughly investigated yet. These two algorithms are described below.

Gibbs Sampler. Assume a multidimensional 2-PL dichotomous model. Let (θ, a, d) denote the collection of ability and item parameters. The Gibbs Sampler simulates data from the joint posterior distribution of (θ, a, d) by drawing from the full conditional distribution, the conditional distribution of one component of the model given the other components in the model (Gelman, 1984). With enough numbers of iterations, the distribution of the simulated data approximates the joint posterior distribution. The issue of convergence is discussed in the next section.

To sample from the full conditional distribution, the adaptive rejection sampling algorithm is used if the distribution is log-concave (Gilks & Wild, 1992). Essentially it constructs an envelope function and a squeezing function, which form upper and lower bounds to the full conditional distribution function. It draws a value from the envelope function and accepts or rejects it as coming from the full conditional distribution depending on whether specific criteria are met. As the sampling proceeds, the envelope function and squeezing function converge to the full conditional distribution function.

For i=1,2,…, n, j=1, 2, …, N and k=1,2,…, m, let Y denote the observed data; $\boldsymbol{\theta}_{jk} \sim iid\ N(0,1)$ represent the ability of person j on the kth dimension; $\boldsymbol{a}_{ik}$ $\sim iid\ N(0,1)I$ (a_{ik} >0) represent the discrimination parameter of item i on the kth dimension, and let $\boldsymbol{\gamma}_{\boldsymbol{i}} \sim iid\ N(0,1)$ represent item difficulty levels. The probability of the

data given (θ, a, γ) is: $L(Y|\,\theta, a, \gamma)=\prod_{i=1}^{m}\prod_{j=1}^{N}\{p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}\}$. (Here p_{ij} refers the same as it in equation (4).)

Let $\pi(\theta)$, $\pi(a)$, $\pi(\gamma)$, denote the prior distributions for $\theta, a,$ and γ. According to equation (6), the joint posterior distribution of (θ, a, γ) given *Y* is:

$$\pi(\theta, a, \gamma|Y) \propto L(Y|\,\theta, a, \gamma)\,\pi(\theta)\,\pi(a)\pi(\gamma)$$

$$\propto L(Y|\,\theta, a, \gamma)\exp\left(-\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{m}\theta_{jk}^{2}-\right.$$

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{m}a_{ik}^{2}$$

$$\left.-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{m}\gamma_{i}^{2}\right)I\,(a_{ik}>0)$$

The full conditional distributions are:

$$\pi(\theta_{jk}|\theta_{jk'}, a, \gamma, k'\neq k) \propto \exp(-\theta_{jk}^{2}/2)\prod_{i=1}^{m}\prod_{j=1}^{N}\{p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}\}$$

$$\pi(a_{ik}|a_{ik'}, \theta, \gamma_i, k'\neq k) \propto \exp(-a_{ik}^{2}/2)\prod_{i=1}^{m}\prod_{j=1}^{N}\{p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}\}\;I(a_{ik}>0)$$

$$\pi(\gamma_i|a_{ik}, \theta) \propto \exp(-\gamma_{i}^{2}/2)\prod_{i=1}^{m}\prod_{j=1}^{N}\{p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}\}$$

Note: For the individual parameters, many of the terms in $\prod_{i=1}^{m}\prod_{j=1}^{N}\{p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}\}$ are

constant relative to that parameter.

Considering arbitrary starting values θ^0, a^0, γ^0, an iteration of the Gibbs Sampler consists of drawing a sample sequentially from the following distributions:

θ_{11}^1 from $\pi(\theta_{11}|\theta_{j'k'}^0, a^0, \gamma^0, j' > 1, k' > 1)$,…,

a_{11}^1 from $\pi(a_{11}|\theta^1, a_{i'k'}^0, \gamma^0, i' > 1, k' > 1)$,…,

and γ_{nm}^1 from $\pi(\gamma_n|\theta^1, a^1, \gamma_{i'}^1, i' < n)$.

Put simply, the Gibbs Sampler draws the new value for each parameter according to its distribution based on the values of all the other parameters in the model. During this process, new values for the parameters are used as soon as they are obtained. For instance, the new value of θ^1 is sampled conditioning on the old values of a^0 and γ^0; the new value of a^1 is sampled conditioning on the new value of θ^1 and the old value of γ^0; the new value of γ^1 is sampled conditioning on the new values of θ^1 and a^1.

HMC-NUTS. Borrowing from Hamiltonian Dynamics in Physics, in which the energy of a system has potential and kinetic parts, for each of the 'position' variables θ_j (referring to the general model parameters rather than person parameter in IRT models), HMC uses a 'momentum' variable φ_j. In HMC, both θ and φ are updated together. The momentum variable φ is an auxiliary variable used to speed up exploration of the parameter space (Gelman, Carlin, Stern, & Rubin, 2014). This feature of HMC allows it to converge to high-dimensional target distributions much faster than simpler methods such as Gibbs Sampler (Hoffman & Gelman, 2014).

The joint posterior distribution in HMC is defined as $p(\theta, \varphi|Y) = p(\varphi)\, p(\theta|Y)$, since $p(\varphi)$ is independent from y. In addition, HMC also requires the gradient of the log-posterior density: $\frac{d\log p(\theta|Y)}{d\theta}$. The variable φ is usually set to have a multivariate normal distribution, with mean of 0 and covariance equal to a prespecified "Mass Matrix": $\boldsymbol{M}$ (so

called by analogy to the physical model of Hamiltonian Dynamics). The Mass Matrix is commonly taken to be a diagonal matrix. (Gelman, Carlin, Stern, & Rubin, 2014).

According to Gelman et al. (2014), HMC proceeds with a series of iterations. There are four steps in one HMC iteration:

Step 1: The iteration begins by randomly drawing φ from its posterior distribution, which is the same as its prior distribution, $\varphi \sim N(0, \boldsymbol{M})$.

Step 2： The main part of HMC is a simultaneous update of (θ, φ) via mimicking of physical dynamics. This update includes *L* leapfrog steps, each scaled by a step size factor ϵ. In each leapfrog step, both θ and φ are updated in relation to the other. The *L* leapfrog steps proceed as follows:

Repeat the steps *L* times:

a) Use the gradient of the log-posterior density of θ to make a half-step of φ:

$$\varphi \leftarrow \varphi + \frac{1}{2}\,\epsilon \frac{d\log p\,(\theta|Y)}{d\theta}.$$

b) Use the vector of φ to update the vector of θ:

$$\theta \leftarrow \theta + \epsilon \mathbf{M}^{-1}\,\varphi$$

Here $\boldsymbol{M}$ is the covariance matrix of φ.

c) Again use the gradient of the log-posterior density of θ to make a half-step of φ:

$$\varphi \leftarrow \varphi + \frac{1}{2}\,\epsilon \frac{d\log p\,(\theta|Y)}{d\theta}.$$

Except in the first and last leapfrog steps, the half-step updates (c) from one leapfrog iteration and (a) from the following iteration can be combined to update φ in a single step. In this view, the updating starts from a half-step of φ, then

performs L-1 full update steps of parameter vector θ and the momentum vector φ, and concludes with a full update of θ and a half-step of φ.

Step 3: Label θ^{t-1}, φ^{t-1}as the value of the position and momentum parameter vectors at the start of the leapfrog process, and θ^*, φ^* as the values after the F steps. In the accept-reject step, we compute: $r=\frac{p\,(\theta^*\,|Y)p\,(\varphi^*)}{p\,(\theta^{t-1}\,|Y)p\,(\varphi^{t-1})}$.

Step 4: Set

$$\theta^t = \begin{cases} \theta^* & with\ probablity \min\ (r, 1) \\ \theta^{t-1} & otherwise. \end{cases}$$

Since we do not care about φ, there is no need to track the updates of it.

HMC can be tuned in three places: the probability distribution for the parameters φ, the scaling factor ϵ, and the number of leapfrog steps per iteration L. For difficult HMC problems, Gelman et al. (2014) suggests the three tuning parameters would be set to vary as the algorithm moves through the posterior distribution, with the diagonal mass matrix $\boldsymbol{M}$ scaling to the local curvature of the log density, the step size getting smaller when the curvature is high, and the number of steps L being large that the algorithm circles around. Specifically, the No-U-Turn Sampler (NUTS) algorithm (Hoffman & Gelman 2014) is an adaptation of HMC, and it determines the number of the leapfrog steps L at each iteration. Essentially, the method extends the trajectory in each iteration until it turns around. This sends the trajectory as far as it can go in that iteration. (In fact, a more elaborate procedure is required to preserve the property that the sequence approached the target distribution.) The full HMC-NUTS also adaptively sets the mass matrix $\boldsymbol{M}$, and the step size ϵ during a warmup phase. The HMC-NUTS algorithm is

complicated and requires more mathematical computations, and the details can be found in Hoffman and Gelman (2014). Fully implemented HMC-NUTS can be accessed via R software through the *rstan* package (Carpenter et al., 2017), which was used in this study.

Summary of Comparative Studies of Estimation Methods in IRT

Test Length. Martin-Fernandez and Revuelta (2017) compared two recent estimation algorithms: Metropolis-Hasting Robbins-Monro (MHRM; Cai, 2010a, 2010b) and HMC-NUTS, with two habitual algorithms: Gibbs Sampler and MML for multidimensional 2-PL dichotomous item response models. In Martin-Fernandez and Revuelta's (2017) simulation study, a test length of 15 was used for the unidimensional model, and a test length of 18 or 25 for the multidimensional model, with the sample size equal to 500 or 1000 for both unidimensional and multidimensional models. The results showed that overall the four estimation methods performed similarly in recovering parameters with less than five latent factors, and HMC-NUTS and MHRM could be regarded as recent improvements over traditional methods: MML and Gibbs Sampler. For estimation accuracy, they found that, as expected, more accurate estimates could be obtained when sample size and test length for each dimension increased, and recovery of intercept parameters was more precise than slope parameters even in poorly defined factors (less than three items) for all the four methods, but the conclusions were not clear regarding which estimation approach performed the best for the recovery of slope parameters in the poorly defined dimension situation. For computing speed, they concluded that MHRM was by far the fastest estimation method, but HMC-NUTS converged faster than the other three in small sample conditions, without clarifying how

small the sample size is. Overall, the performance of these methods for parameter recovery in smaller samples and shorter test lengths remains unexplored.

Prior Choice. Natesan et al. (2016) studied the impact of vague, matched, and hierarchical priors using Gibbs Sampler and Variational Bayesian estimation methods in unidimensional 1-PL and 2-PL dichotomous models. Overall, hierarchical prior and matched priors performed better, and the vague priors produced large errors or convergence issues in parameter recovery using both algorithms, which are not recommended. The authors recommended hierarchical priors considering estimation accuracy and time effectiveness.

In multidimensional 2-PL dichotomous models, Martin-Fernandez and Revuelta (2017) also concluded that for low and high informative priors, the differences in recovery of model parameters were negligible using HMC-NUTS, the differences were small using Gibbs Sampler, and the differences were significant when using MML with small samples. The authors also mentioned that the differences would be more prominent in real applications since their study had 500 simulees, a sample size that was larger than most of the application studies. The literature only focused on the dichotomous models, and there remain concerns when generalizing their results to multidimensional and polytomous models.

Interdimensional Correlation. Kuo and Sheng (2016) compared several parameter estimation methods: two MML approaches (Bock-Aitkin expectation-maximum algorithm, adaptive quadrature approach), four fully Bayesian algorithms (Gibbs sampling, Metropolis-Hastings, Hastings-within-Gibbs, blocked Metropolis), and

the MHRM algorithm for multi-unidimensional (two dimensions) 2-PL graded response IRT models via different statistical software packages. The authors concluded that when the correlation between the two dimensions was low ($\rho = .20$), these estimation methods provided similar results. However, if the two dimensions were moderately or highly correlated ($\rho > .50$), Hastings-within-Gibbs, one of the fully Bayesian algorithms, recovered the discrimination and the inter-dimension correlation parameters better. One of the limitations of this study was not including the relatively new and efficient parameter estimation method: HMC-NUTS in its Bayesian algorithm category. Furthermore, discussion of the number of dimensions was limited to two, and they limited the polytomous model to only three categories. Hence, further studies are needed to evaluate the estimation methods for polytomous models with more than three categories.

Purpose of the Study

The purpose of this study was to compare the performance of two Bayesian MCMC estimation approaches: Gibbs Sampler and HMC-NUTS in M2PPC models under different simulated conditions. The goal was to demonstrate how these estimation methods perform in M2PPC models in a set of realistic conditions including variations in test length, prior choice, and interdimensional correlation. Monte Carlo simulation was used to generate data. The accuracy of the parameter estimation was evaluated by Root Mean Squared Errors (RMSEs). In addition, four-way analyses of variance (ANOVAs) of the Root Mean Squared Errors (RMSEs) and Bias for each of the three parameter recovery parameters (θ, a, γ) in the M2PPC model were conducted, and the effect sizes

ω^2 were calculated to assess the effects of independent variables and interactions. Additionally, computational speed for the two algorithms was recorded for comparisons. Therefore, the specific research questions of this study are as follows:

1. Which of the two estimation approaches (Gibbs Sampler and HMC-NUTS) yields the more accurate estimates in M2PPC model?
2. Do test length, interdimensional correlation, and prior choice influence the accuracy of the two estimation approaches in M2PPC model?
3. On balance, considering the computational speed and estimation accuracy, which of the two approaches performs better in parameter recovery for an M2PPC model?

Researchers and practitioners can utilize the results of this study in making informed decisions on research design and scale construction when using M2PPC model in their data collection and analysis.

Delimitations

- This study only focuses on the 2PL IRT model. More parameters such as in 3PL models are not considered.
- This study only focuses on two Bayesian MCMC estimation methods. Since the recent analytical MHRM (Cai, 2010a, 2010b) estimation approach has already shown better performance than traditional MML, future studies can incorporate MHRM into comparisons.
- This study only examines simulated data.

- The simulation design manipulates several factors including sample size, test length, number of choice categories, the number of dimensions, interdimensional correlation, and prior choice, which limits the generalizability of the results of this simulation study as other factors are fixed.
- All the latent traits in this study followe a normal distribution. It is possible that in behavioral and medical scales, the person ability parameter follows a skewed distribution. Therefore, the robustness of the parameter recovery with normal priors OR estimation with skewed priors needs to be investigated in future studies.

Definition of Terms

Item response theory (IRT)

IRT is the theory used in educational and psychological measurement that investigates a mathematical relationship between persons' abilities and item responses.

Multidimensional IRT (MIRT)

MIRT assumes multiple traits are measured by each item. For instance, in psychological settings, an item for screening depression may also measure patient anxiety levels.

Partial Credit Model (PCM) and Generalized Partial Credit Model (GPCM)

PCM is a unidimensional model for the analysis of responses recorded in two or more ordered categories. GPCM was formulated by Muraki (1992) based on Masters' (1982) partial credit model (PCM) by relaxing the assumption of uniform discriminating power of test items.

Multidimensional 2 Parameter Partial Credit Model (M2PPC)

M2PPC is a multidimensional version of GPCM model with slopes and intercepts to be estimated, which is the focus of this study.

Monte Carlo (MC) Simulation

MC simulation is used to describe a process for propagating uncertainties in model inputs into uncertainties in model outputs (results). It is a kind of simulation that quantitatively represents uncertainties. MC simulation methods are set up as experiment, in which data are generated to test hypotheses theoretically derived (Paxton, Curran, Bollen, Kirby, & Chen, 2001).

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.

Prior Probability Distribution

In Bayesian statistical inference, a prior probability distribution, often called the prior, of an uncertain quantity is the probability distribution that expresses one's belief about this quantity before some evidence accumulated (Gelman et. al, 2014).

Posterior Probability Distribution

In Bayesian statistics, the posterior probability of an uncertain proposition is the conditional probability that is assigned after the relevant evidence is considered. Similarly, the posterior probability distribution is the probability distribution of an

unknown quantity, conditional on the evidence obtained from an experiment (Gelman et. al, 2014).

Root Mean Square Error (*RMSE*)

The *RMSE* is a measure of the accuracy of parameter estimates, which measures the average squared discrepancy between a set of estimated and true parameters, which can be conceived as the amount of variability around a point estimate. In this study, the log transformation of RMSE (log*RMSE*) was used to meet the normality assumption of Analysis of Variance (ANOVA).

Bias

Bias is another measure of the accuracy of parameter estimates. In this study, the log transformation of Bias (log*Bias*) was used to meet the normality assumption of ANOVA.

Computational Speed

Computational speed is the time it takes an estimator to get the parameter estimates, which is recorded for estimator comparison.

Jags

Jags stands for just another Gibbs Sampler, which is a program for analysis of Bayesian models with MCMC simulation using Gibbs sampling.

Stan

Stan is a computer program which implements Bayesian inference using HMC-NUTS.

Chapter Two: Method

Introduction

Chapter Two includes a detailed description of the study's methodology. First, I describe the Monte Carlo (MC) simulation design. Then, I explain the three phases of my analysis, which were data generation, parameter estimation, and parameter recovery evaluation. The data generation phase includes generation of response datasets of different test lengths, and interdimensional correlations, in the multidimensional two parameter partial credit (M2PPC) model. Scenarios were created to illustrate different conditions in the datasets. Datasets with different conditions were estimated using different priors with two Bayesian approaches: Gibbs Sampler and Hamiltonian Monte Carlo-No-U-Turn-Sampler (HMC-NUTS). Finally, performance of the two Bayesian estimation approaches under different conditions was evaluated using parameter recovery criteria.

Monte Carlo Simulation

This study involved a Monte Carlo (MC) simulation. MC simulation is an empirical method for generating datasets for evaluating the performance of statistical models. MC simulation methods are employed as experiments, where data are generated to test theoretically derived hypotheses (Paxton, Curran, Bollen, Kirby, & Chen, 2001). In MC simulation, each input parameter is defined by a source distribution, from which

random samples are drawn. MC simulation allows researchers to evaluate the finite sampling performance of estimators by creating manipulated conditions, where sampling distributions of parameter estimates are produced. Sampling distributions are theoretical, but MC methods can be used to create simulated data reflecting the characteristics of sampling distributions. In this study, I drew samples of parameters ($\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{\gamma}$) from specific distributions to simulate M2PPC model response data, see equation (5). Subsequently, three Bayesian estimation methods with different priors were used to compute estimates on simulated datasets to address my research questions.

Analysis Procedures

The analysis in this study included three phases: (1) Data generation, (2) Parameter estimation, (3) Outcome analysis. There are multiple steps beneath each of the three phases, and the following is a detailed description of the steps.

Phase I Data Generation

To compare the aforementioned two Bayesian methods in estimating the M2PPC model, three factors were controlled before simulating response data: test length (n), and interdimensional correlation (ρ). The choice of n, and ρ was based on previous research using similar models. When investigating multidimensional graded response models (GRMs), the simulation study in Fu, Tao, and Shi (2010) used $N($sample size$)$=1000, n=20 and ρ=0.2, 0.4,…0.8 for polytomous items involving three categories. Sheng and Wikle (2008) adopted $N($sample size$)$=1000, n=18, ρ=0.2, 0.5,…0.8 in their simulation studies with dichotomous multi-dimensional models. Working with nominal response

models, Wollack, Bolt, Cohen, and Lee (2002) found that the parameter recovery was improved by increasing the test length from 10 to 30 items in their simulation studies.

Manipulated Variables. In the current study, three variables were manipulated: (1) test length: n= 15 and 30; (2) sample size: N=500; (3) interdimensional correlation: ρ= 0.2, 0.5, and 0.8. The two levels of test length were chosen to simulate different measuring instruments in psychological and behavioral settings, such as personality inventories. Based on the previous simulation studies (e.g., Fu et al., 2010; Sheng & Wike, 2008), and the complexity of the M2PPC model structure that can elongate MCMC estimation time for larger sample size unreasonably, N= 500 was chosen as the sample size, and three levels of interdimensional correlations were adopted to simulate low, medium, and high correlations between the three dimensions. The three factors were fully crossed, resulting in, 2 (n)* 3 (ρ) =6 simulation conditions.

Based on the literature (e.g., Revilla, Saris, & Krosnick, 2014; Weijters, Cabooter, & Schillewaert, 2010) about the number of response categories, a 5-point scale is preferred for psychological and behavioral tests when the respondents are the general public. Therefore, a three-dimensional 2-parameter partial credit model with five response categories was used in this study. There are four steps for generating data: (1) generating person ability parameter values (θs), (2) generating intercept parameter values (γs), (3) generating item discriminating parameter values (as), and (4) simulating the M2PPC model response data.

Persons. A person latent ability vector $\boldsymbol{\theta}$ was generated from a multivariate normal distribution with a mean vector of 0s, and a covariance matrix with 1s along the

diagonal. The off-diagonal elements represented the correlations of any two of the marginal distributions, which were specified above (ρ= 0.2, 0.5, and 0.8).

Items. The intercept parameter vector $\boldsymbol{\gamma}$ was generated from a standard normal distribution. The item discriminating parameter vector $\boldsymbol{a}$ was generated from a lognormal distribution: μ=0 and σ^2=0.25, with the expected value of 1.13 and a variance of 0.36. The person and item parameters were used in model (5) to simulate the response data in R software, *mirt* package (R. Philip Chalmers, 2012).

Replications. To increase the generalizability of the results, each of the conditions needs to be replicated a number of times. However, there are inconsistencies in the literature about the number of replications: Jiang et al. (2016) employed 30 replications when investigating the sample size requirements for estimation of a multidimensional graded response model; Natesan et al. (2016) chose to replicate 100 times to examine the Bayesian prior choice unidimensional IRT model using MCMC; Kuo and Sheng (2016) used 10 replications for each of their conditions for comparison of estimation methods in a multi-unidimensional graded response model with sample size equals to 500 or 1000. In this study, following Kuo and Sheng, I generated 10 datasets for each of the conditions. Although according to Harwell, Stone, Hsu, and Kirisci (1996), a minimum of 25 replications should be carried out for typical IRT-based MCMC studies, this study carried out 10 replications due to the computational expense of the MCMC algorithms for test conditions such as *N*=500 and *n*=30.

Phase II Parameter Estimation with Different Priors

In the Bayesian estimation framework, setting different priors allows for systematic incorporation of previous information into the current parameter estimation. Even though the influence of prior information on estimation decreases as the sample size increases, the selection of priors still impacts the estimation when the sample size is small (Levy & Mislevy, 2016). Therefore, the choice of priors is a crucial issue when using a Bayesian framework to estimate parameters in IRT models. However, the Bayesian IRT literature is inconsistent on the choice of priors. For example, Swaminathan and Gifford (1982) discouraged the employment of extremely optimistic priors such as the unit/standard normal distribution, or diffused priors with large variances, but Sheng and Headrick (2012) still used informative normal (i.e., $\mu_a = \mu_\gamma = 0, \sigma_a^2 = \sigma_\gamma^2 = 1$) or uniform priors for a_i and γ_i . Moreover, these optimistic priors and diffused priors continue to be used in Bayesian frameworks for IRT models, such as Fox (2010). Until recently, there was no guidance on what kind of priors need to be chosen for different IRT models. Natesan et al. (2016) investigated the impact of vague, matched, and hierarchical priors in using two Bayesian estimation methods in unidimensional 1-PL and 2-PL dichotomous models, and suggested future research needs to be done for more complicated IRT models. Building upon the literature, this study investigated the choice of vague, matched, and hierarchical priors in M2PPC models using two Bayesian estimation methods: Gibbs Sampler and HMC-NUTS.

Matched Prior. Matched prior refers to the distribution that was used to generate response data, which is an unrealistic situation. However, this was used as the reference prior with which other prior results were compared. For each person, the ability

parameter was set to: $\theta \sim N\ (\mathbf{0}, \sum)$, $\sum$ defined by $var\theta = 1$, $cov\ (\theta_k, \theta_m)$=$\rho$ for low inter-dimensional correlation ρ =0.2, medium inter-dimension correlation ρ =0.5, and high inter-dimension correlation ρ =0.8; For each item, each response intercept parameter was set to: $\gamma \sim N\ (0,1)$; For each dimension, each item discrimination parameter was set to: $a \sim lognormal\ (0, 0.5)$

Standard Vague Prior. Vague prior refers to a large uncertainty about the prior distribution of the parameters. The uncertainty is reflected in the variance of the prior distribution. In the current study, the prior distribution for person ability parameters θ was set to the same as it was in the matched prior. The prior distribution for the discrimination parameter for each dimension and each item was set to $a \sim$ *lognormal* (0, 8) and the intercept parameter prior for each item and each response was set to $\gamma \sim N\ (0, 10^3)$, and the large standard deviations: 8 and 10^3 represented the degree of uncertainty.

Hierarchical Prior. Hierarchical prior refers to the parameters of the prior distribution are regarded as random variables and are given vague hyper-priors. For each person, the prior for the ability parameter was set to: $\theta \sim N\ (\boldsymbol{m_\theta}, \sum_\theta)$, $\boldsymbol{m_\theta} \sim N\ (0, 10^6)$, and $\sum_\theta$~Inv-Wishart($\sum_{\theta 0}$,4); for each item and each response the prior for the intercept parameter was set to: $\gamma \sim N\ (m_\gamma, u_\gamma^{-1})$, $m_\gamma \sim N\ (0, 10^6)$, and u_γ~gamma (1, 1); for each dimension and each item the prior for the discrimination parameter was set to: $a \sim$*lognormal* (m_a, u_a^{-1}), $m_a \sim N\ (0, 10^6)$, and u_a~gamma (1, 1). An informative inverse gamma (1,1) distribution was used for variance because Gelman & Rubin (1992) cautioned against the use of low values such as 0.001 for gamma priors leading to improper posteriors.

Gibbs sampler and HMC-NUTS were implemented for each simulated data set using *rjags* and *rstan* (Carpenter et al., 2017), respectively, where the burn-in stage was set to 5,000 iterations followed by three chains with 11,000 iterations. For both algorithms, the initial values for the discrimination parameters (*as*) for the 3 chains were set to follow the *lognormal* distribution, and those for the intercept parameters (γs) and person ability parameters (θs) were set to follow standard normal distribution. The convergence of Markov chains was evaluated using the Gelman-Rubin R statistic (Gelman & Rubin, 1992). To get the Gelman-Rubin R statistic, several Markov chains are generated with spread initial points from the parameter space. Then, the Gelman-Rubin R statistic can be calculated by comparing the variance within and between chains. Suppose ξ is the parameter of interest. Further, suppose *M* Markov chains were generated, each with a length of *H* after initial draws are thrown away (e.g., *H*=burnin=5000). Denote ξ_{im} as the simulated parameter in the *i*th generation of the *m*th chain. The between chain variance is defined as $B = \frac{H}{M-1}\sum_{m=1}^{M}(\overline{\xi_{.m}} - \overline{\xi_{..}})^2$, and the within chain variance is defined as $W = \frac{1}{M}\sum_{m=1}^{M} S_g^2$, where $S_g^2 = \frac{1}{H-1}\sum_{i=1}^{N}(\overline{\xi_{im}} - \overline{\xi_{.m}})^2$. An Gelman-Rubin R statistic is obtained as $\hat{R} = \sqrt{\frac{\widehat{var}(\xi|\mathbf{y})}{W}}$, where $\widehat{var}(\xi|\mathbf{y}) = \frac{H-1}{H}W + \frac{1}{H}B$. Brooks and Gelman (1998) noted that $\hat{R} < 1.20$ provides evidence that the chain has converged to the posterior distribution. If the $\hat{R}$ is larger than 1.20, it suggests that the Markov chains have not reached stationarity and more iterations are needed to improve convergence.

Phase III Outcome Analysis: Parameter Recovery Evaluation Criteria

Root Mean Squared Error. The root mean squared error (RMSE) measures the average squared discrepancy between a set of estimated and true parameters and can be regarded as the amount of variability around a point estimate. The vector of RMSE for the intercept parameter for a test length *n* was computed as:

$$\text{RMSE } (\gamma) = \frac{\sum_{r=1}^{S}\sqrt{\frac{1}{n}\sum_{i=1}^{n}||\boldsymbol{\gamma_{ir}}-\widehat{\boldsymbol{\gamma_{ir}}}||^2}}{S}, \quad \text{for } i=1, 2, \ldots, n \text{ and } r=1,2,\ldots S \tag{7}$$

where $\widehat{\boldsymbol{\gamma_{ir}}}$ and $\boldsymbol{\gamma_{ir}}$ were the estimated and real values of the intercept parameter vector for replication *S* and item *n*. The vector of RMSE for the discrimination parameter for a test length *n* was computed as:

$$\text{RMSE } (a) = \frac{\sum_{r=1}^{S}\sqrt{\frac{1}{n}\sum_{i=1}^{n}||\boldsymbol{a_{ir}}-\widehat{\boldsymbol{a_{ir}}}||^2}}{S}, \text{ for } i=1, 2, \ldots, n \text{ and } r=1,2,\ldots S \tag{8}$$

where $\widehat{\boldsymbol{a_{ir}}}$ and $\boldsymbol{a_{ir}}$ were the estimated and real values of the discrimination parameter vector for replication *S* and item *n*. The vector of RMSE for the ability parameter for a given sample of *N* persons was calculated as:

$$\text{RMSE } (\theta) = \frac{\sum_{r=1}^{S}\sqrt{\frac{1}{N}\sum_{j=1}^{N}||\boldsymbol{\theta_{jr}}-\widehat{\boldsymbol{\theta_{jr}}}||^2}}{S}, \quad \text{for } j=1, 2, \ldots, N \text{ and } r=1,2,\ldots S \tag{9}$$

where $\widehat{\boldsymbol{\theta_{jr}}}$ and $\boldsymbol{\theta_{jr}}$ were the estimated and real values of the person ability parameter vector for replication *S* (*S* is the total number of replications, which equals 10 in this study) and person *N*.

Finally, RMSEs are examined based on comparison rather than some absolute cutoffs. Therefore, the lower the RMSE value is, the more accurate the estimate is (Natesan et al. 2016).

Bias. Also, the accuracy of parameter estimation was evaluated using average of Bias.

The vector of average Bias of the intercept parameter for a test length *n* was computed as:

$$Bias\ (\gamma) = \frac{\sum_{r=1}^{S}(\frac{1}{n}\sum_{i=1}^{n}||\widehat{\boldsymbol{\gamma}_{ir}}||-||\boldsymbol{\gamma}_{ir}||)}{S},\ \text{for } i=1, 2, \ldots, n \text{ and } r=1,2,\ldots S \quad (10)$$

where $\widehat{\boldsymbol{\gamma}_{ir}}$ and $\boldsymbol{\gamma}_{ir}$ were the estimated and real values of the intercept parameter vector for replication *S* and item *n*. The vector of average Bias of the discrimination parameter for a test length *n* was computed as:

$$Bias\ (a) = \frac{\sum_{r=1}^{S}(\frac{1}{n}\sum_{i=1}^{n}||\widehat{\boldsymbol{a}_{ir}}||-||\boldsymbol{a}_{ir}||)}{S},\ \text{for } i=1, 2, \ldots, n \text{ and } r=1,2,\ldots S \quad (11)$$

where $\widehat{\boldsymbol{a}_{ir}}$ and $\boldsymbol{a}_{ir}$ were the estimated and real values of the discrimination parameter vector for replication *S* and item *n*. The vector of average Bias of the ability parameter for a given sample of *N* persons was calculated as:

$$Bias\ (\theta) = \frac{\sum_{r=1}^{S}(\frac{1}{N}\sum_{j=1}^{N}||\widehat{\boldsymbol{\theta}_{jr}}||-||\boldsymbol{\theta}_{jr}||)}{S},\ \text{for } j=1, 2, \ldots, N \text{ and } r=1,2,\ldots S \quad (12)$$

where $\widehat{\boldsymbol{\theta}_{jr}}$ and $\boldsymbol{\theta}_{jr}$ were the estimated and real values of the person ability parameter vector for replication *S* and person *N*. If *Bias* (θ) is close to zero, it suggests that the value of the estimated person ability parameter is close to the true person ability parameter. If *Bias* (θ) is positive, it suggests that the person with higher ability is estimated with even higher ability. If *Bias* (θ) is negative, it suggests that the person with higher ability is estimated with relative lower ability.

RMSE and Bias for each of the parameter in M2PPC model were calculated based on 10 replications for each of the fully crossed conditions (2 test lengths, 3 inter-

dimension correlations, 3 priors, and 2 estimation methods), which results in 36 conditions (See Table 1).

Table 1

Simulation Design

Test Length	Inter-dimension Correlation	Priors	Estimation Methods
15	0.2	Standard	Gibbs Sampler
		Vague	HMC-NUTS
		Matched	Gibbs Sampler
			HMC-NUTS
		Hierarchical	Gibbs Sampler
			HMC-NUTS
	0.5	Standard	Gibbs Sampler
		Vague	HMC-NUTS
		Matched	Gibbs Sampler
			HMC-NUTS
		Hierarchical	Gibbs Sampler
			HMC-NUTS
	0.8	Standard	Gibbs Sampler
		Vague	HMC-NUTS

		Matched	Gibbs Sampler
			HMC-NUTS
		Hierarchical	Gibbs Sampler
			HMC-NUTS
30	0.2	Standard	Gibbs Sampler
		Vague	HMC-NUTS
		Matched	Gibbs Sampler
			HMC-NUTS
		Hierarchical	Gibbs Sampler
			HMC-NUTS
	0.5	Standard	Gibbs Sampler
		Vague	HMC-NUTS
		Matched	Gibbs Sampler
			HMC-NUTS
		Hierarchical	Gibbs Sampler
			HMC-NUTS
	0.8	Standard	Gibbs Sampler
		Vague	HMC-NUTS
		Matched	Gibbs Sampler
			HMC-NUTS

Hierarchical	Gibbs Sampler	
	HMC-NUTS	

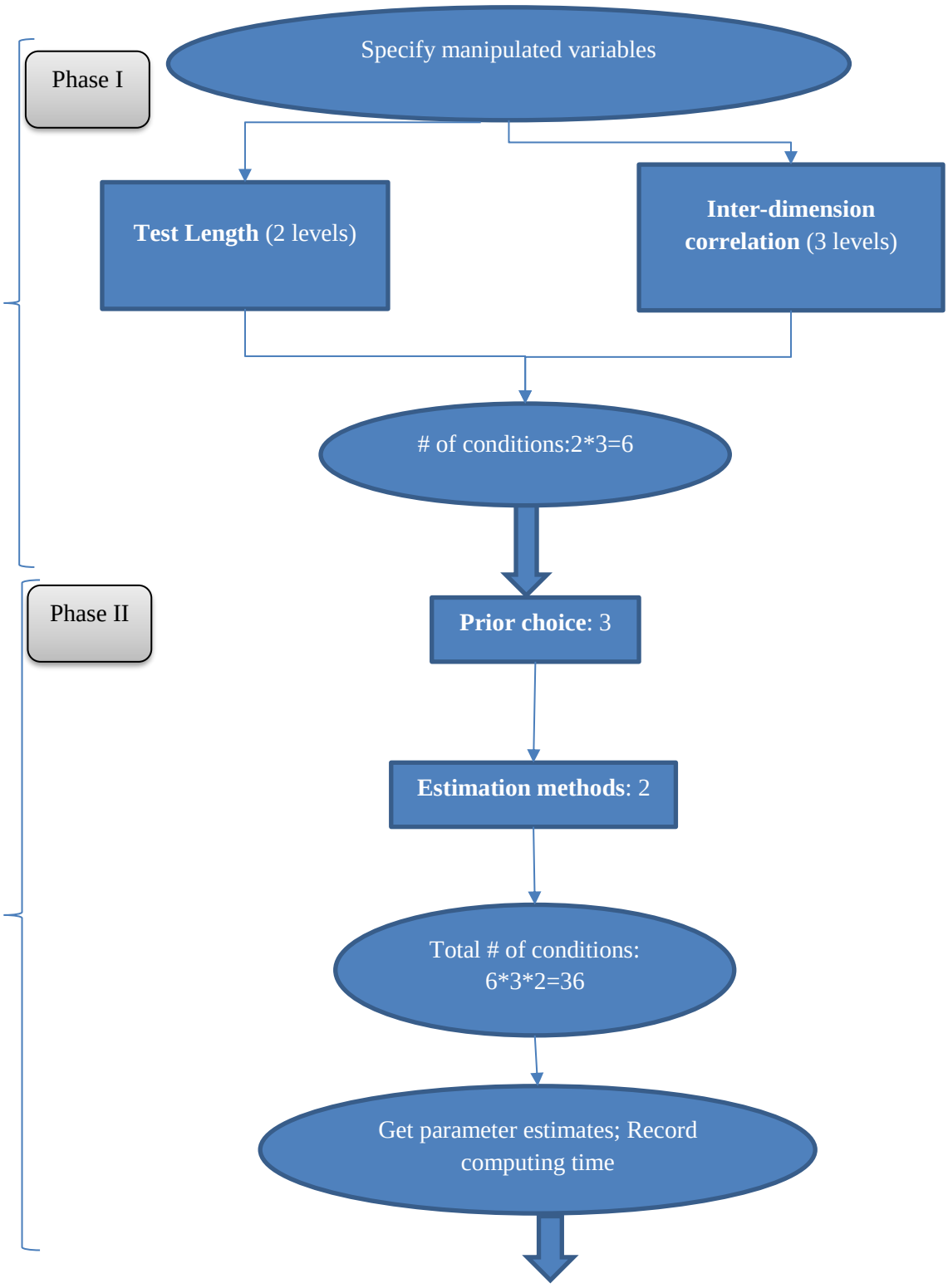
Four-way Analysis of Variance. In order to examine which factor or interactions accounted for the most variance in the estimation accuracy of the parameters in M2PPC model with the respect to test length, interdimensional correlation, prior choice, estimation method, and all possible interactions, 20 four-way analyses of variances (ANOVAs) were conducted. The 20 four-way ANOVAs were carried out on the 20 dependent variables: log*RMSE* of person ability estimates on three dimensions, intercept estimates for the four transitional points, and discrimination estimates on three dimensions, and log*Bias* of person ability estimates on three dimensions, intercept estimates for the four transitional points, and discrimination estimates on three dimensions. For all the four-way ANOVAs, the same independent variables were used: test length (two levels), inter-dimension correlation (three levels), prior choice (three levels), estimation method (two levels), estimation method (two levels), and all possible interactions. Log transformation was used for the dependent variables in the sets of ANOVA analyses. Log transformation was conducted in order to meet one of the ANOVA assumptions-- normal distribution of the dependent variable.

Effect sizes (ω^2) for each recovered parameter were calculated to assess the effects of each independent variable and possible interactions, which was defined as:

$$\omega^2 = \frac{SS_{effect} - (df_{effect})(MS_{error})}{SS_{total} + MS_{error}}$$

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 squared error, and SS_{total} is the total sum of squares of the ANOVA model. ω^2 was evaluated according to Cohen (1988): large effect size is greater than 14%, medium effect size is greater than 6%, and small effect size is greater than 1% of the total variance. The reason the ANOVA summary table with p-values were not included in the result section of this study is because when the effect sizes are exactly zero or very small, even if the p-values are significant, the differences are often meaningless. Moreover, in this particular study, the major purpose is to examine the substantive effects of each of the independent variable on accunting for the variance of the dependent variables rather than the statistical significance.

Apart from the above evaluation criteria, computational time for each recovered parameter was also recorded for comparison. Figure 1 is a visual representation of the analysis procedure.


Phase I
Specify manipulated variables
Test Length (2 levels)
Inter-dimension correlation (3 levels)
# of conditions:2*3=6
Phase II
Prior choice: 3
Estimation methods: 2
Total # of conditions: 6*3*2=36
Get parameter estimates; Record computing time

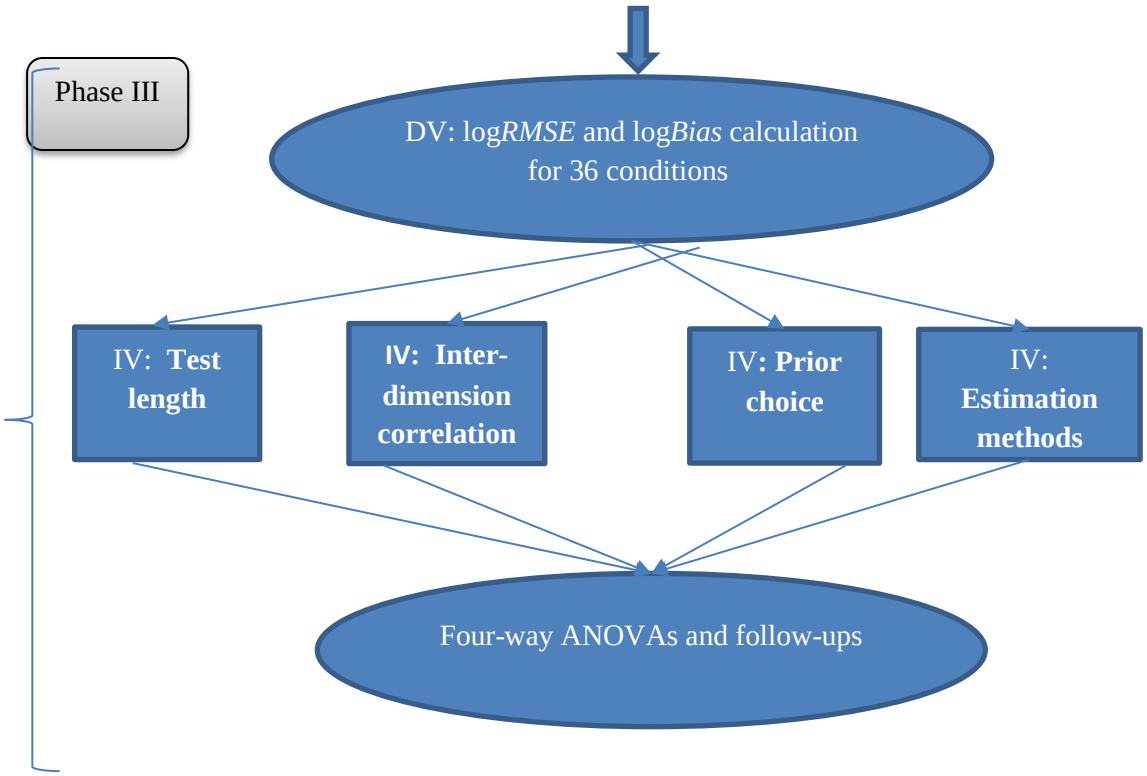


Figure 1. Visual Representation of Analysis Procedure

Chapter Three: Results

Introduction

Chapter Three summarizes the results of parameter estimation in a multidimensional two parameter partial credit (M2PPC) model using different priors with different test lengths and interdimensional correlations conditions with Gibbs Sampler and Hamiltonian Monte Carlo-No-U-Turn-Sampler (HMC-NUTS). First, the convergence of the Markov chains is described respectively for Gibbs Sampler and HMC-NUTS. Second, the two indices: root mean squared error (RMSE) and Bias are summarized for item and person parameters for different conditions (described thoroughly in Chapter Two) in the M2PPC model. Also, I include the estimation speed for different conditions using Gibbs Sampler and HMC-NUTS. Last, 20 four-way ANOVAs [test length (2) x inter-dimension correlation (3) x prior choice (3) x estimation method (2)] were carried out on the 20 dependent variables: log*RMSE* for four intercepts for the four transitional points, three slopes for three dimensions and three person ability estimates for three dimensions; and log*Bias* for four intercepts for the four transitional points, three slopes for three dimensions and three person ability estimates for three dimensions. The chapter concludes with a summary of the findings.

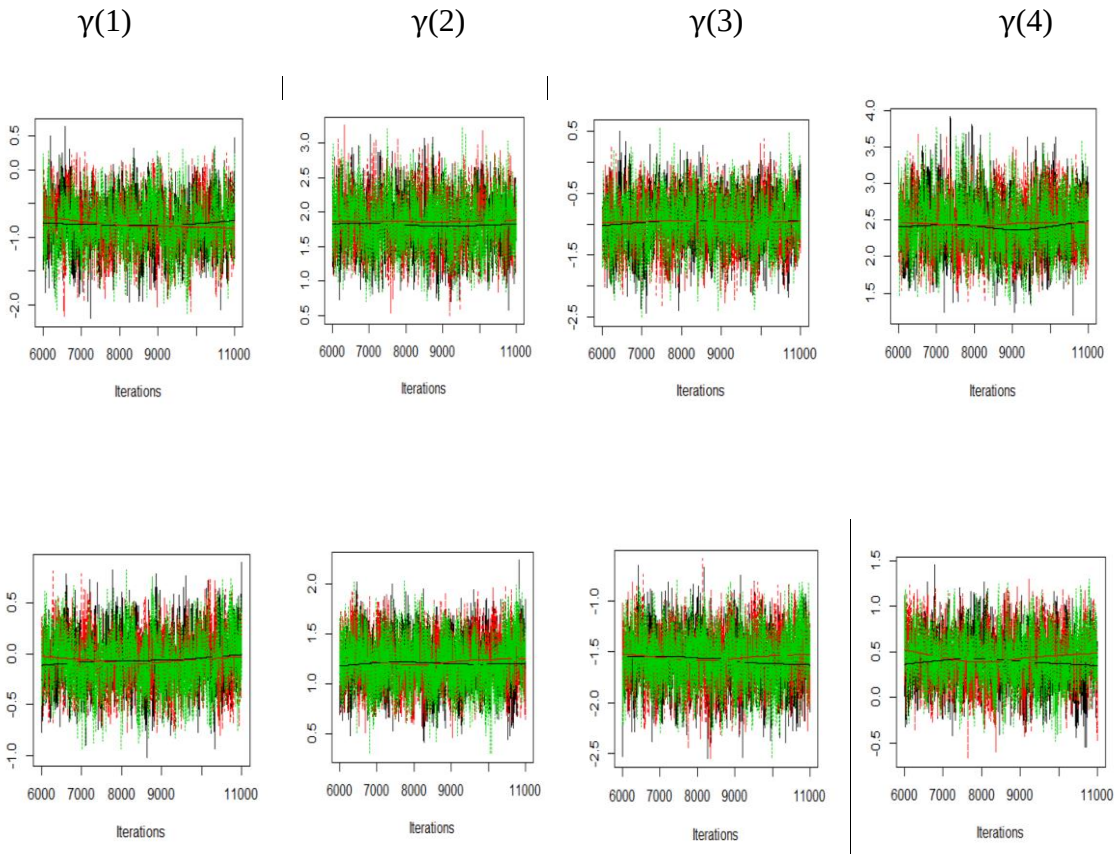
Convergence of Markov Chains

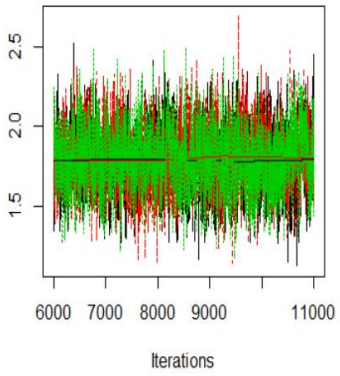
The convergence of Markov chains was evaluated using the Gelman-Rubin R statistic (Gelman & Rubin, 1992), which was detailed in Chapter Two. Moreover, trace plots and Gelman-Rubin plots were generated to help visually check the convergence of the chains. For the M2PPC model estimation, the burn-in stage for Gibbs Sampler (or warm-up for HMC-NUTS) was set to 5,000 followed by three chains with 11,000 iterations, which was tuned by the author. For each simulated dataset, there were 10 estimated parameters: intercept estimates for the four transitional points, discrimination estimates on three dimensions, and person ability estimates on three dimensions. $\hat{R}$s were all less than 1.20 for each of the 10 M2PPC model parameters for all the 36 conditions using both Gibbs Sampler and HMC-NUTS, which suggests that the Markov chains converged to the stationary posterior distribution for all the estimated parameters (Gelman & Rubin, 1992).

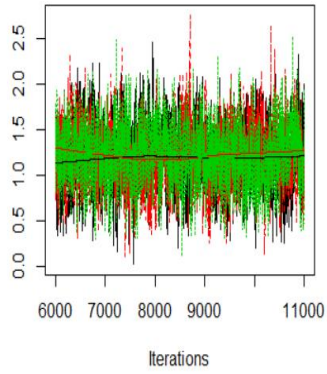
Trace Plots

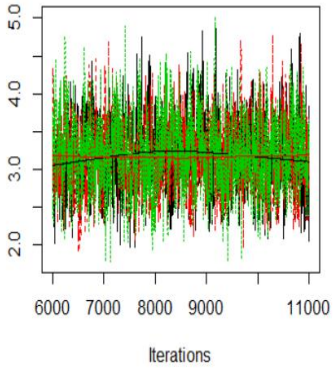
A trace plot displays a plot of iterations vs. sampled values for each estimated parameter in the chain, and it provides an important tool for assessing mixing of the chain. In the trace plots, we want to avoid situations where the chain stays in the same state for too long or for too many consecutive steps in one direction. In this case, trace plots were generated by randomly choosing one item and one case from each dataset falling into one of the 36 different conditions, as shown in Figure 2. The three colors in the each of the trace plots represent three Markov chains. With the illustrated item and person, the trace plots of intercept estimates for the four transitional points,

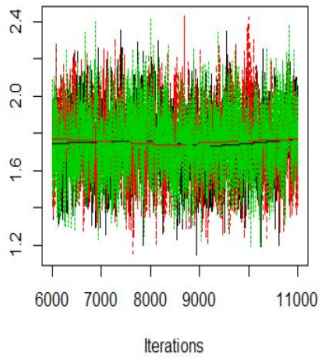
discrimination estimates on three dimensions, and person ability estimates on three dimensions using Gibbs Sampler and HMC-NUTS estimation methods do not demonstrate any orphaned chains for either item or person ability parameters, which suggests the chains mixed well and converged to a stationary distribution (see Figure 2).

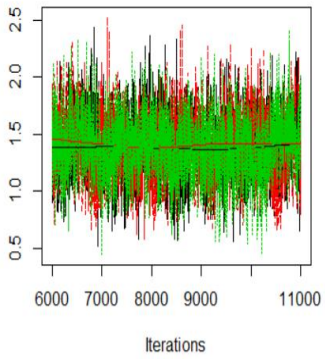


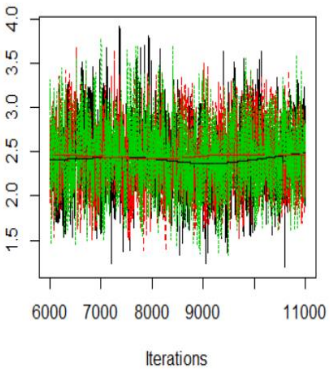




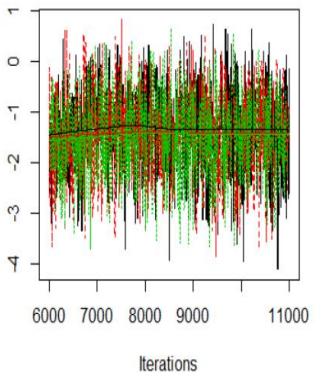


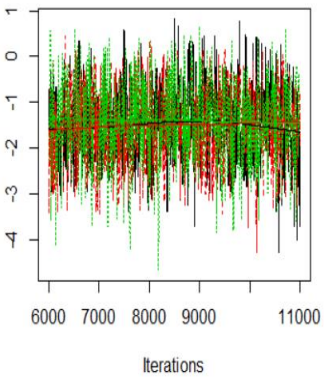


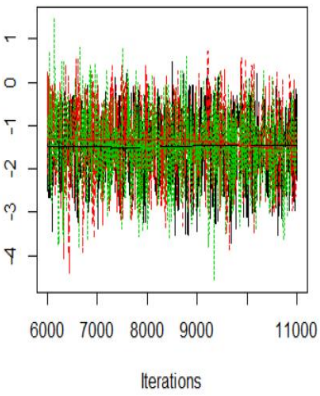




$\theta(1)$ $\theta(2)$ $\theta(3)$







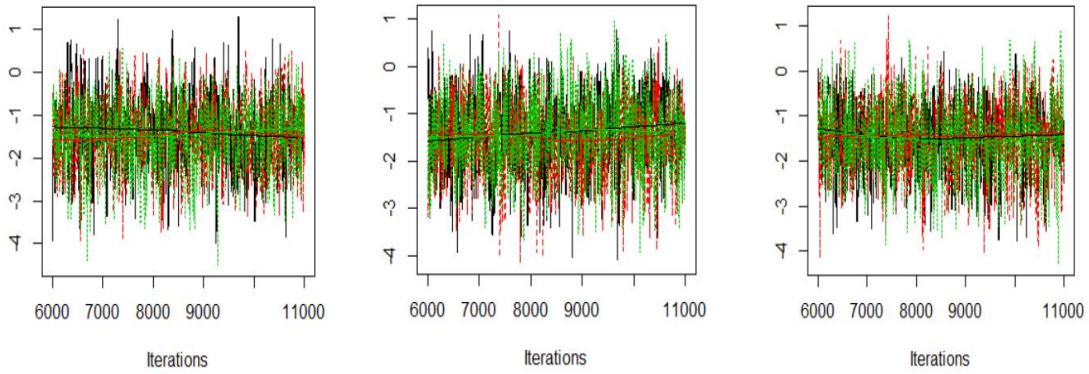
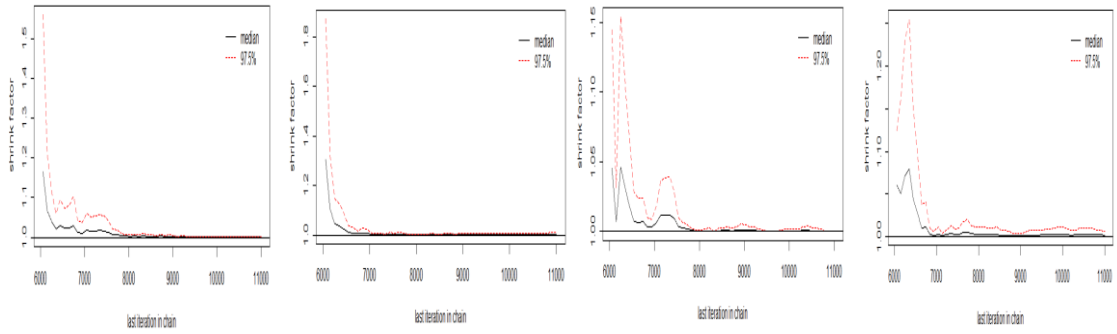


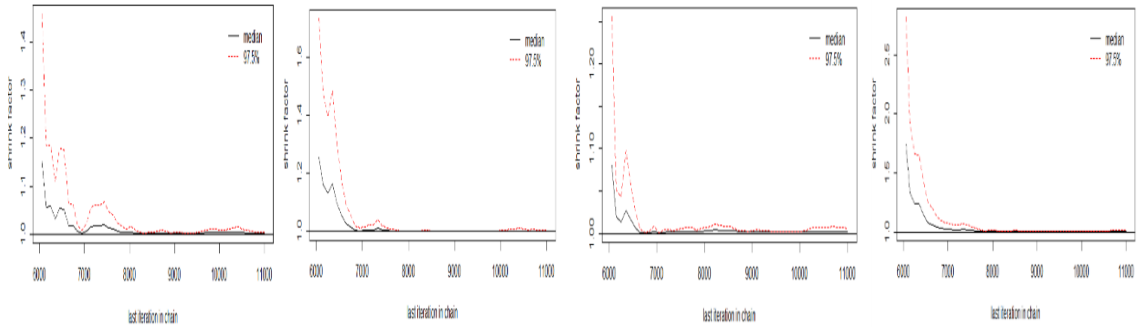
Figure 2. Trace Plots of the Intercept (γ(1), γ(2), γ(3) and γ(4)), Slope Parameter (a(1), a(2), and a(3)), and Person Ability Parameter (θ(1), θ(2), and θ(3)) in the M2PPC Model Using Gibbs Sampler (upper panel) and HMC-NUTS (lower panel).

Gelman-Rubin Plots

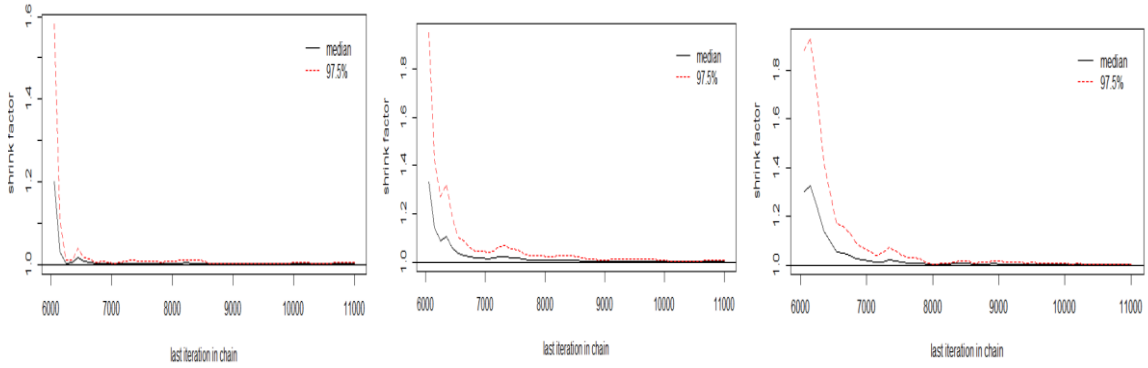
A Gelman-Rubin plot shows the evolution of Gelman-Rubin's shrink factor (Gelman-Rubin statistic: $\hat{R}$) as the number of iterations increases. In the Gelman-Rubin plots, we want to observe the value of Gelman-Rubin's shrink factor approaching one as the number of iterations increases. In these plots, the chains initially are different, but after 6,000-8,000 iterations, they mix together around one in the sample space (the baseline in each plot), as shown in Figure 3. With the illustrated item and person, the Gelman-Rubin plots of intercept estimates for the four transitional points, discrimination estimates on three dimensions, and person ability estimates on three dimensions using Gibbs Sampler and HMC-NUTS estimation methods suggest the chains reached convergence (see Figure 3).

$\gamma(1)$ $\gamma(2)$ $\gamma(3)$ $\gamma(4)$





$a(1)$ $a(2)$ $a(3)$



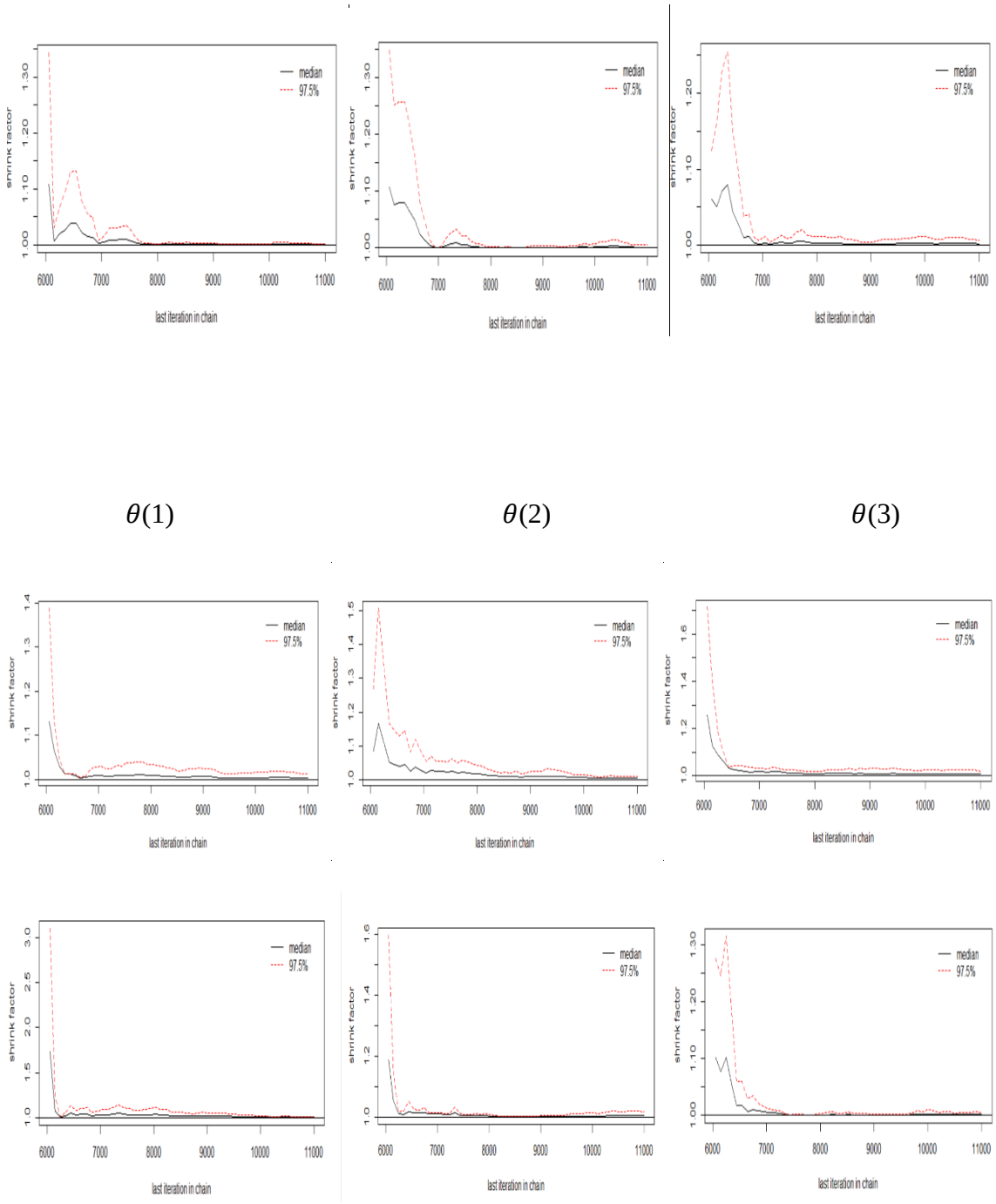


Figure 3. Gelman-Rubin of the Intercept (γ(1), γ(2), γ(3) and γ(4)), Slope Parameter (a(1), a(2), and a(3)), and Person Ability Parameter (θ(1), θ(2), and θ(3)) in the M2PPC Model Using Gibbs Sampler (upper panel) and HMC-NUTS (lower panel).

Item Parameter Recovery

The RMSE and Bias values were averaged across items for evaluation of the recovery of intercept estimates for the four transitional points (intercepts): γ(1), γ(2), γ(3), and γ(4) and discrimination estimates (slopes) on three dimensions: a(1), a(2), and a(3) in the M2PPC model using Gibbs Sampler and HMC-NUTS, and are summarized in Tables 2 through 5.

In terms of estimating γs, the visual representations of intercept *RMSE*s: $\gamma(1)$, $\gamma(2)$, $\gamma(3)$, and $\gamma(4)$ are summarized in Figures 4 through 7. For estimating a, the visual representations of estimation of slope RMSEs: *a(1)*, *a(2)*, and *a(3)* are summarized in Figures 8 through 10. By examination of the tables and figures, Gibbs Samplers and HMC-NUTs did not differ to any great degree in their RMSEs and Bias. Both algorithms recovered the γs and *a*s with similar precision as the *RMSE*s and *Bias* are nearly identical.

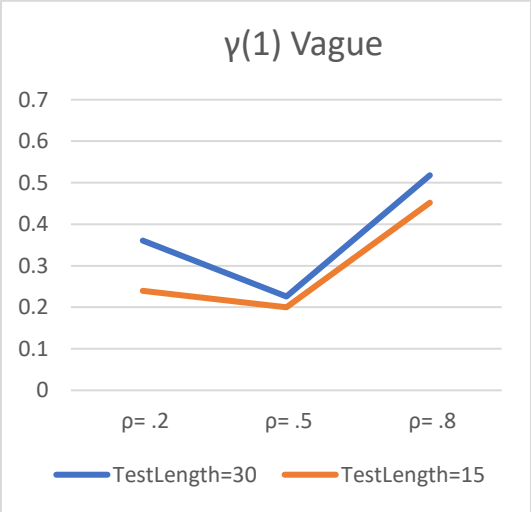
A consistent pattern of the *RMSE*s for both γs and *a*s was found. (From Figure 4 to Figure 10, the orange lines representing test length =15 items are all below the blue lines representing test length = 30 items.) *RMSE*s for both parameters decreased with a decrease in test length regardless of priors and interdimensional correlations, which means the accuracy of the recovery of γs and *a*s increased as the test length decreased.

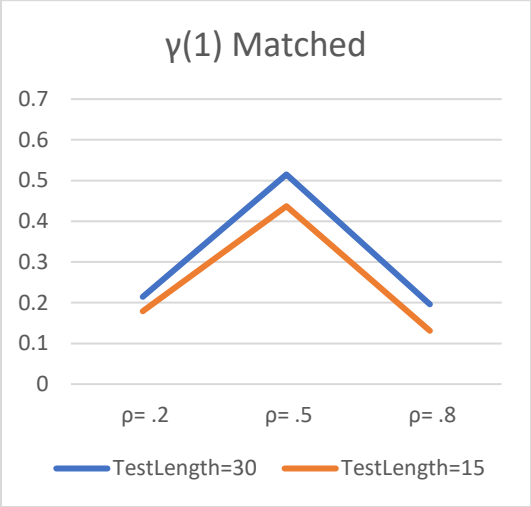
Regarding the prior choice, Matched priors and Hierarchical priors recovered both γs and *a*s better than the Vague priors, since overall the *RMSE*s and absolute values of Bias are relatively larger when using Vague priors than using Matched priors and

Hierarchical priors regardless of test length and interdimensional correlations. Furthermore, for test length=30, the Hierarchical priors recovered $\gamma(1)$ better than Matched priors with lower *RMSE*s; the Matched priors recovered $\gamma(2)$ better than Hierarchical priors with lower *RMSE*s; the recovery of $\gamma(3)$ and $\gamma(4)$ were almost the same (with similar *RMSE*s) using either Matched priors or the Hierarchical priors. However, for test length=30, the recovery of *a*s showed a consistent pattern using different priors. Hierarchical priors always yielded smaller *RMSE*s than Matched priors for both algorithms in all three different interdimensional correlation conditions.

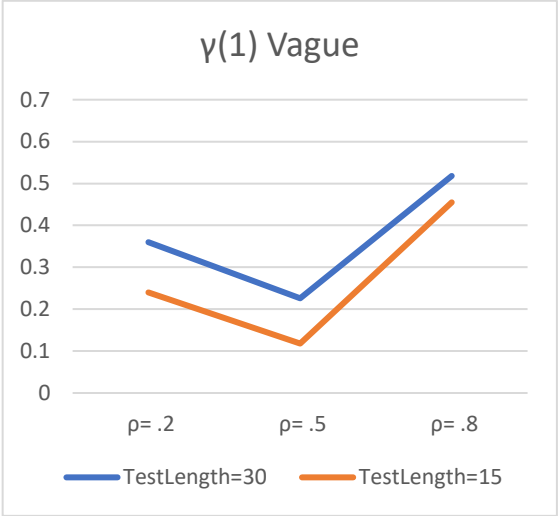
In addition, both algorithms recovered the slope parameters (*a*s) better than the intercept parameters (γs) in all the conditions taking both *RMSE*s and *Bia*s values into consideration (see Tables 2 through 5). There was no consistent trend for the influences of interdimensional correlation on both γs and *a*s recovery considering the test length, two algorithms and the three prior choices. Moreover, there were inconsistencies in the direction of Bias for both γs and *a*s.

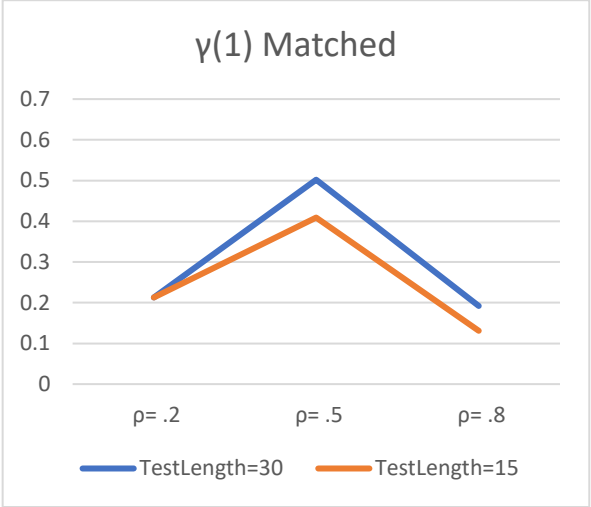
Gibbs Sampler

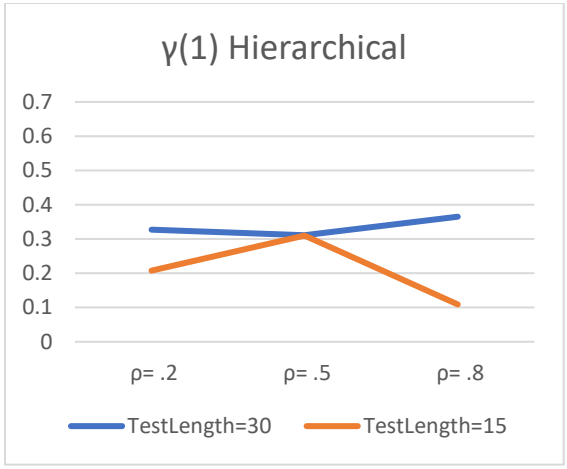




HMC-NUTS







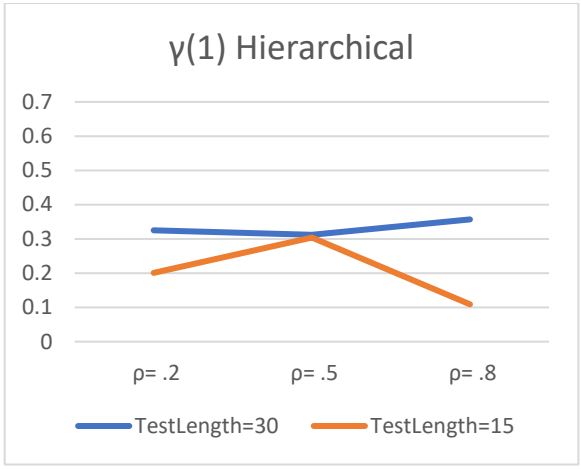
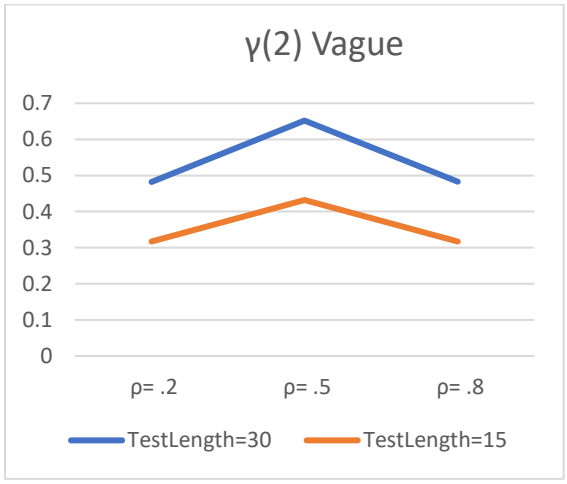
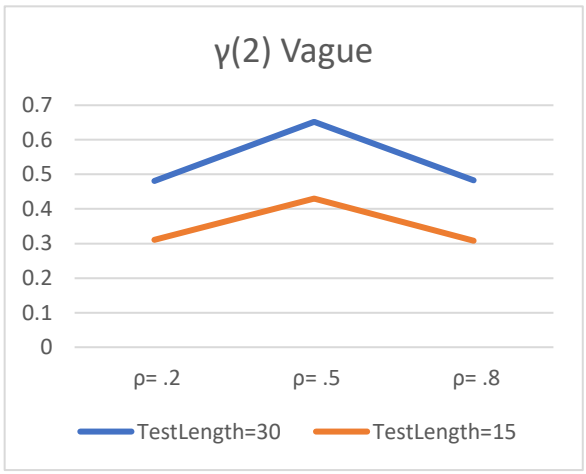


Figure 4. Average RMSE for Recovering Intercept Parameter γ(1) Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC Model.

Gibbs Sampler HMC-NUTS





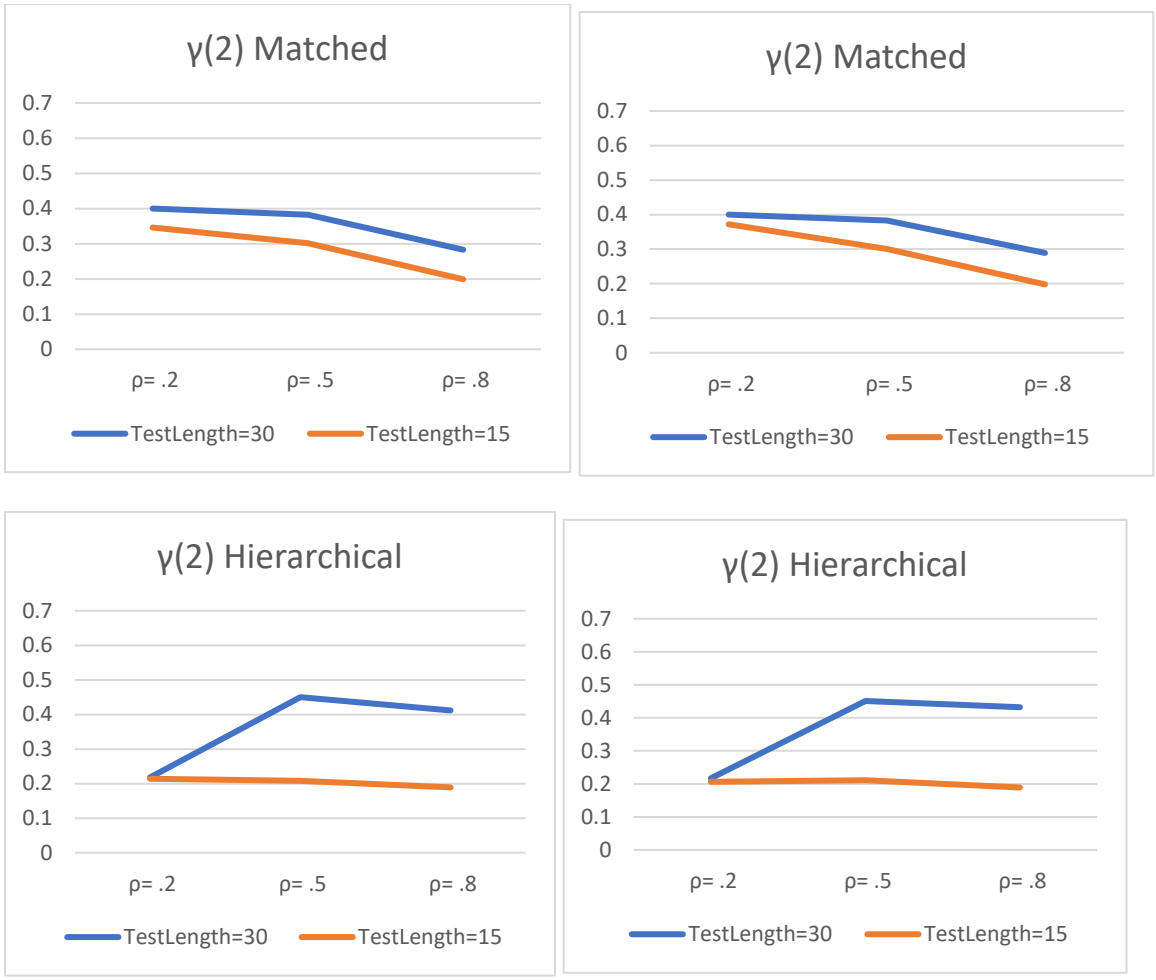


Figure 5. Average RMSE for Recovering Intercept Parameter γ(2) Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC Model.

Gibbs Sampler HMC-NUTS

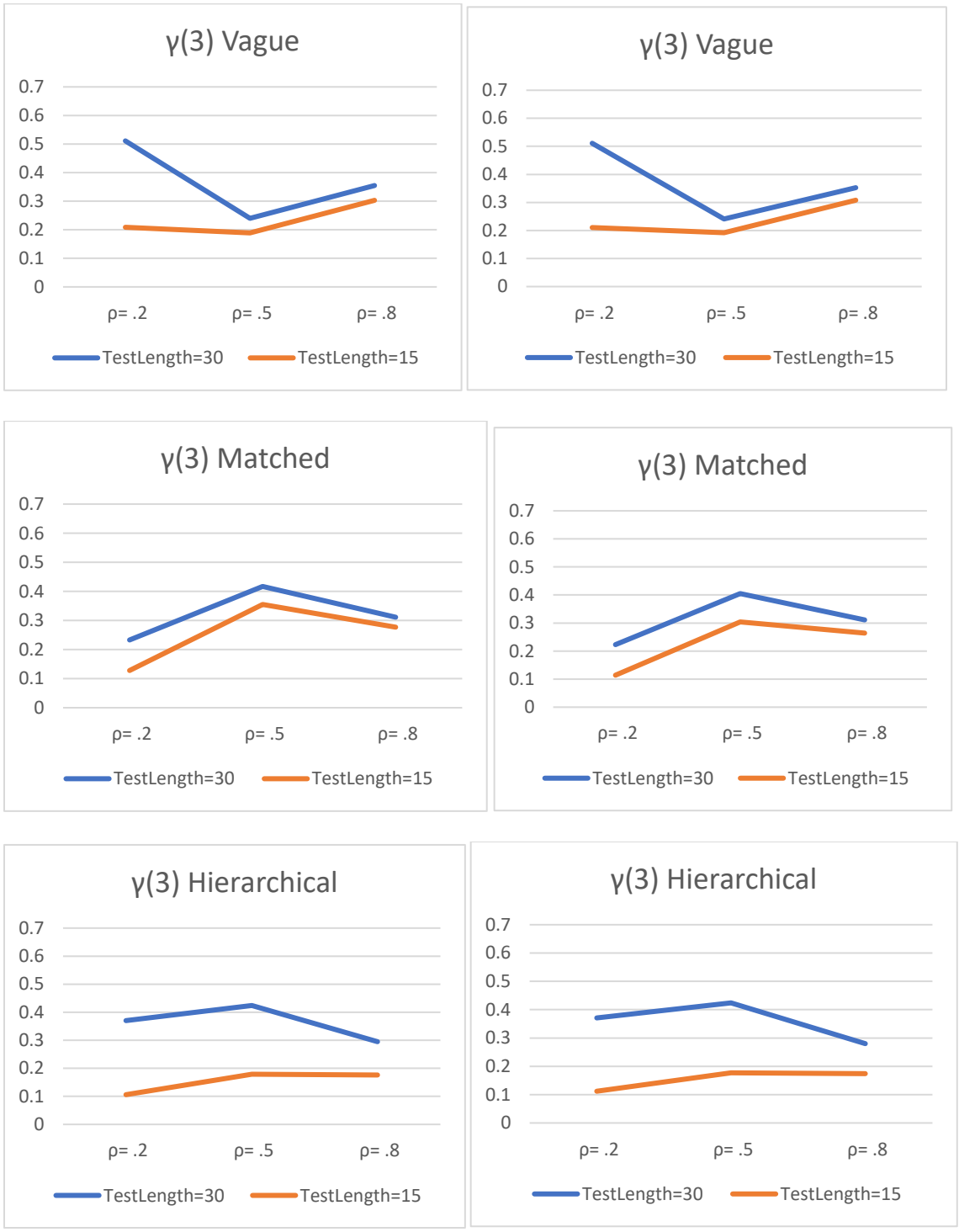
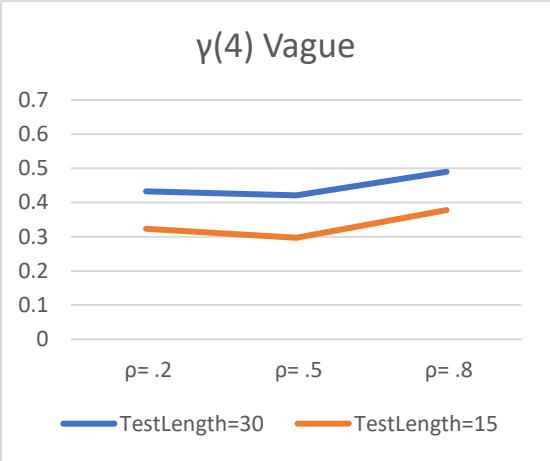
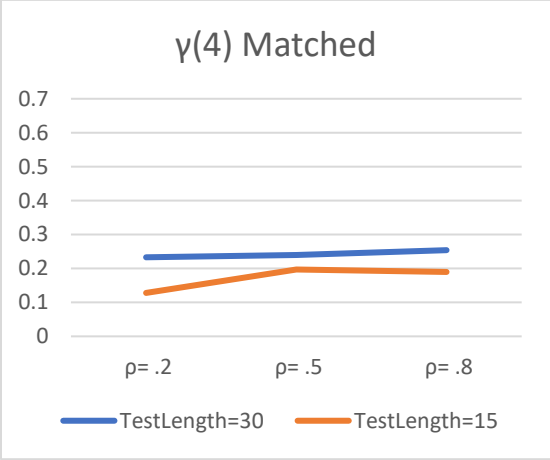
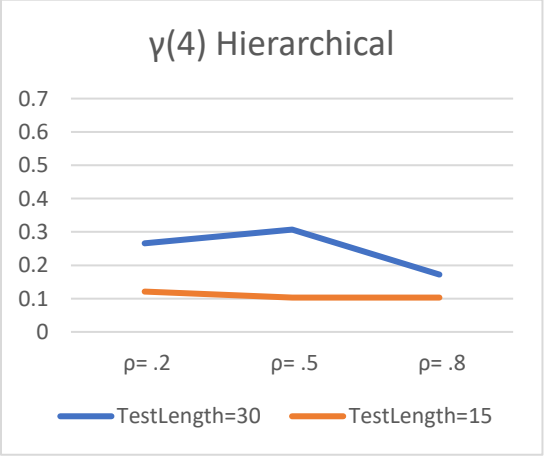


Figure 6. Average RMSE for Recovering Intercept Parameter γ(3) Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC Model.

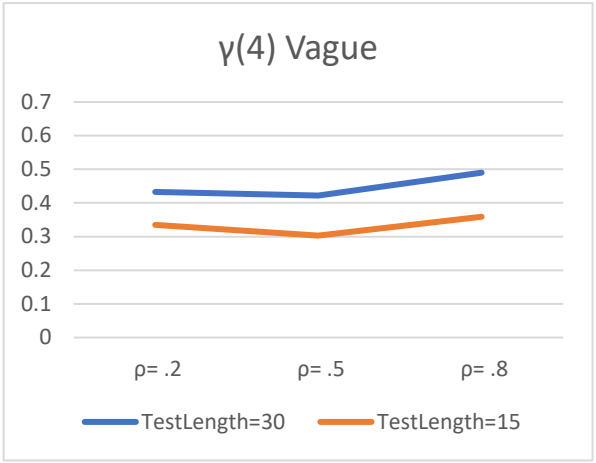
Gibbs Sampler

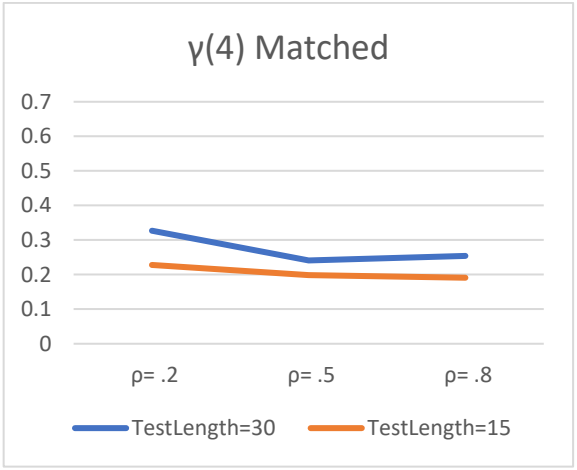






HMC-NUTS





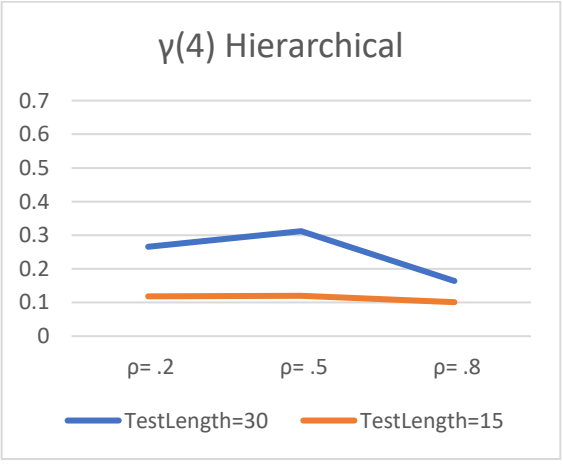
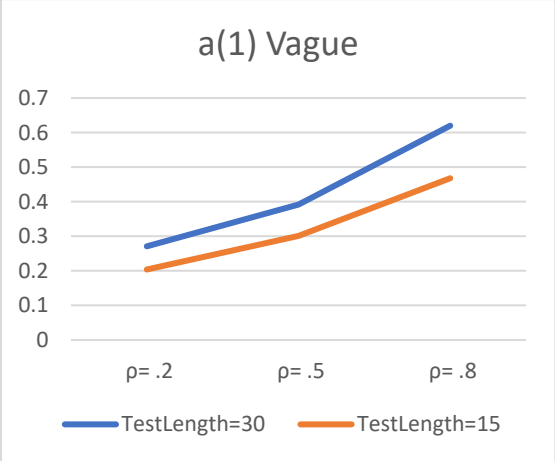
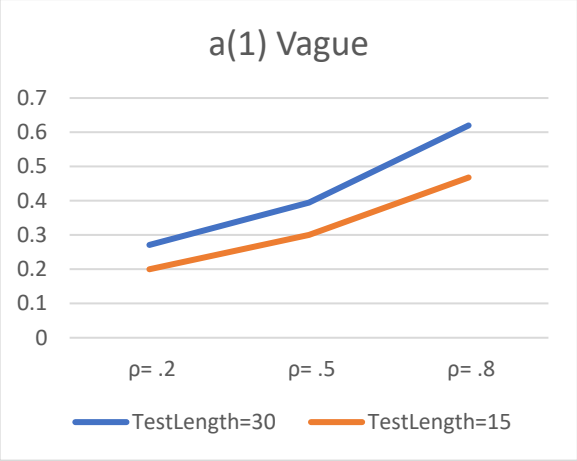


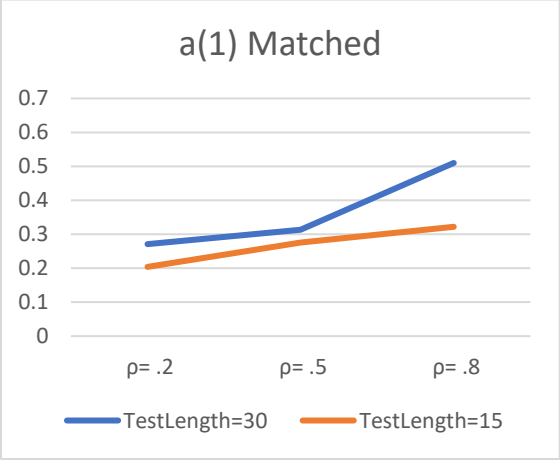
Figure 7. Average RMSE for Recovering Intercept Parameter γ(4) Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC Model.

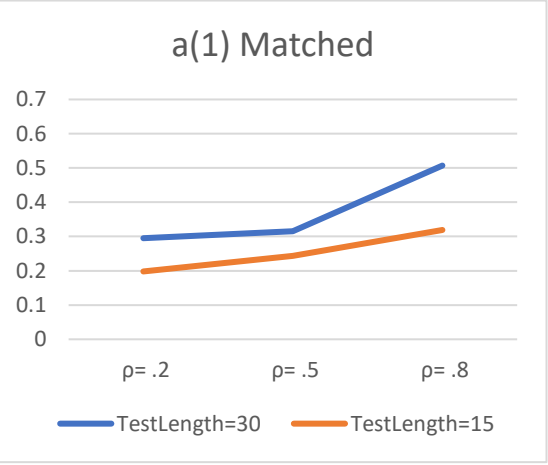
Gibbs Sampler

HMC-NUTS









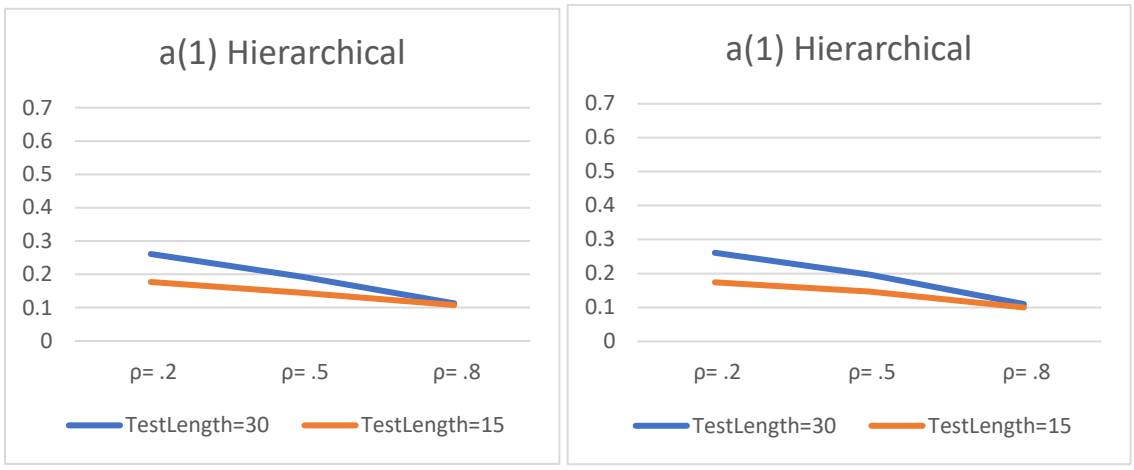
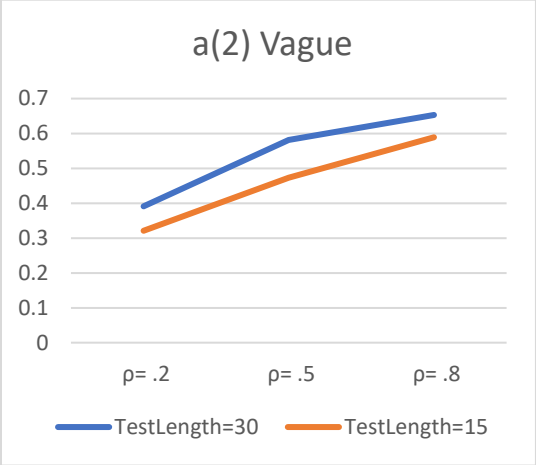
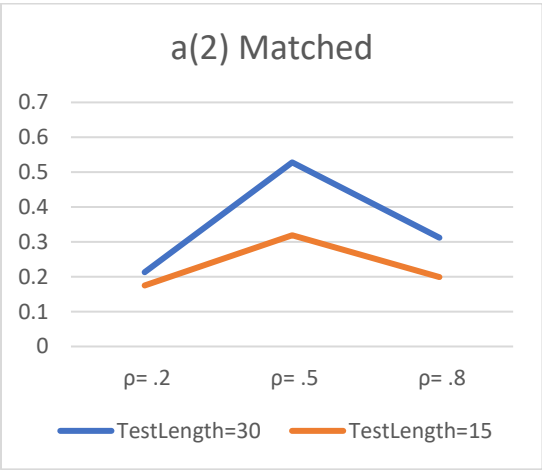
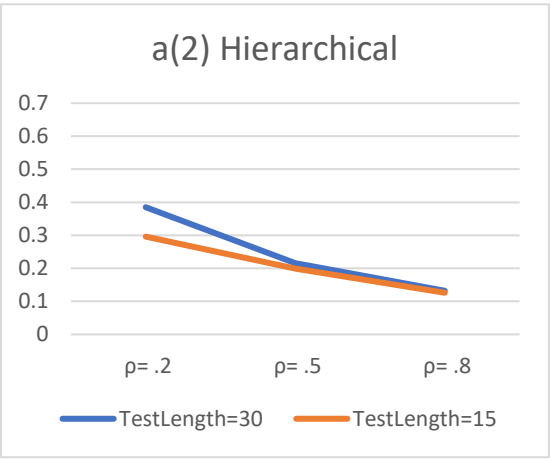


Figure 8. Average RMSE for Recovering Slope Parameter *a(1)* Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC Model.

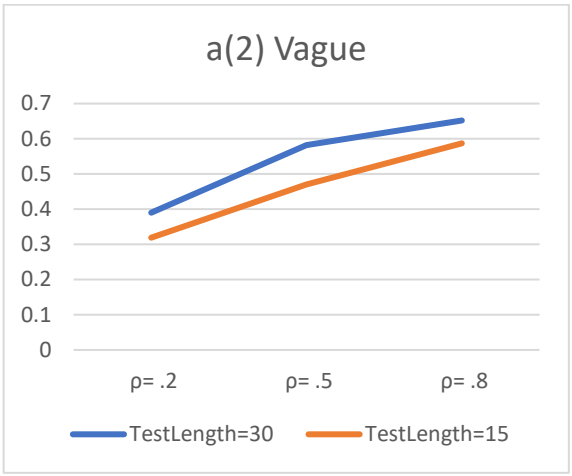
Gibbs Sampler

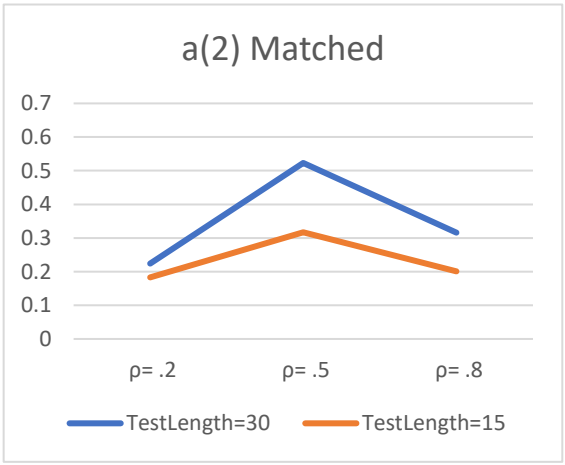






HMC-NUTS





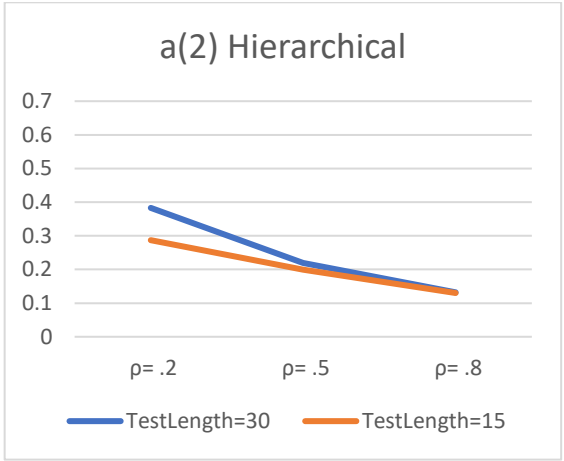


Figure 9. Average RMSE for Recovering Slope Parameter *F* Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC Model.

Gibbs Sampler

HMC-NUTS

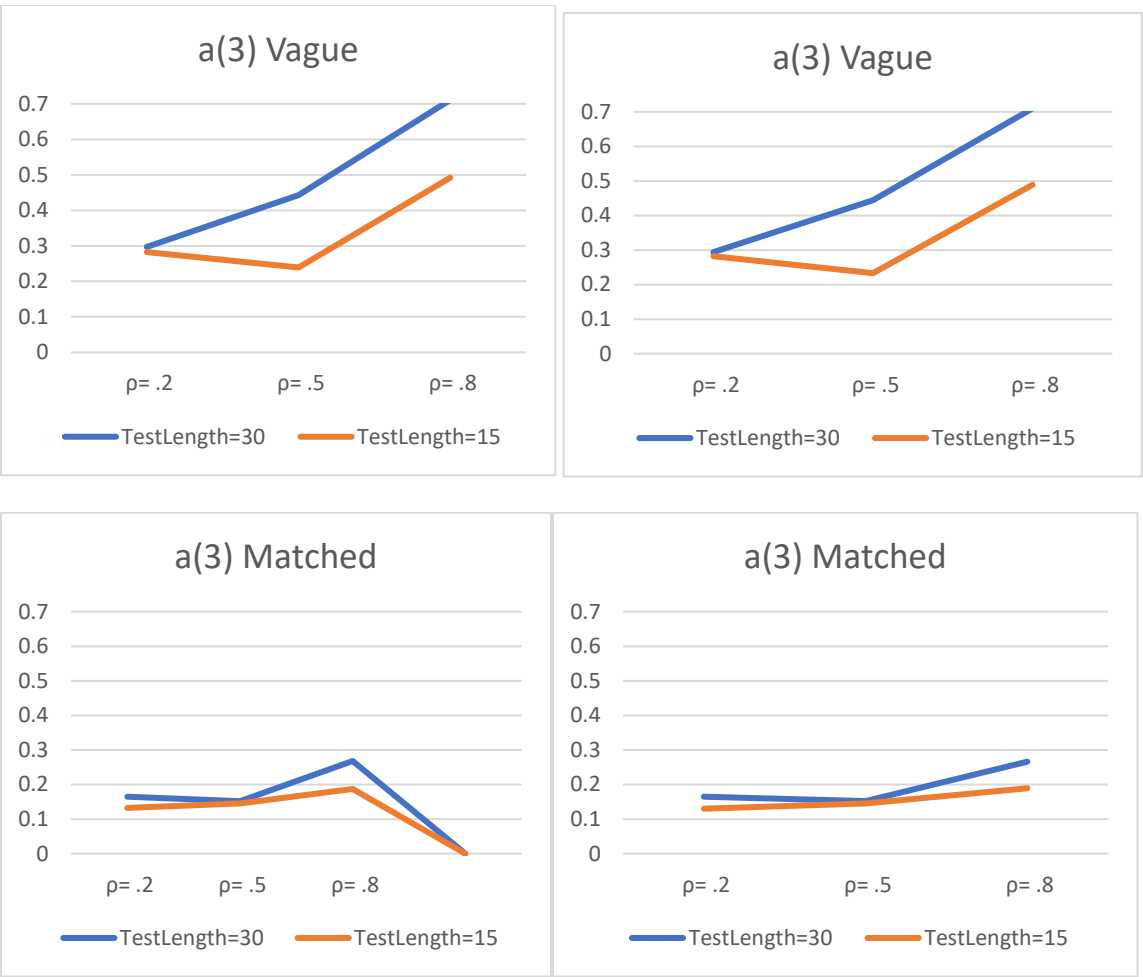


Figure 10. Average RMSE for Recovering Slope Parameter *a(3)* Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC Model.

Table 2

RMSE and Bias of Intercept Estimates for the Four Transitional Points: γ(1), γ(2), γ(3), and γ(4) in the M2PPC Model When Test Length = 30

Prior	Inter-dimension Correlation	Parameters	Gibbs Sampler		HMC-NUTS	
			RMSE	Bias	RMSE	Bias
Vague	.2	γ(1)	.361	-.021	.360	-.021
		γ(2)	.482	-.029	.481	-.031
		γ(3)	.511	-.062	.511	-.062
		γ(4)	.433	-.033	.433	-.032

	.5	γ(1)	.226	-.026	.226	-.025
		γ(2)	.652	-.053	.652	-.055
		γ(3)	.240	-.030	.241	-.030
		γ(4)	.421	.004	.422	.004
	.8	γ(1)	.518	.011	.518	.010
		γ(2)	.483	-.020	.483	-.021
		γ(3)	.355	-.044	.353	-.045
		γ(4)	.490	.010	.490	.010
Matched	.2	γ(1)	.214	-.002	.213	-.002
		γ(2)	.400	-.015	.400	-.014
		γ(3)	.233	-.017	.223	-.017
		γ(4)	.310	-.030	.327	-.031
	.5	γ(1)	.515	.014	.502	.012
		γ(2)	.382	.004	.383	.004
		γ(3)	.417	.040	.405	.041
		γ(4)	.240	-.012	.241	-.011
	.8	γ(1)	.196	-.007	.192	-.007
		γ(2)	.283	-.014	.289	-.013
		γ(3)	.311	-.027	.311	-.027
		γ(4)	.254	-.001	.254	-.001
Hierarchical	.2	γ(1)	.327	-.003	.325	-.002

		γ(2)	.218	-.001	.217	-.001
		γ(3)	.370	-.001	.371	-.002
		γ(4)	.266	-.071	.266	-.071
	.5	γ(1)	.311	-.004	.312	-.004
		γ(2)	.450	-.001	.451	-.001
		γ(3)	.424	-.001	.424	-.002
		γ(4)	.307	.001	.312	.001
	.8	γ(1)	.365	-.011	.357	-.011
		γ(2)	.412	-.001	.432	-.001
		γ(3)	.295	-.003	.280	-.002
		γ(4)	.172	-.014	.164	-.011

Table 3

RMSE and Bias of Slope Estimates on the Three Dimensions: a(1), a(2), and a(3) in the M2PPC Model When Test Length = 30

Prior	Inter-dimension Correlation	Parameters	Gibbs Sampler		HMC-NUTS	
			RMSE	Bias	RMSE	Bias
Vague	.2	*a*(1)	.271	-.102	.271	-.101
		a(2)	.391	-.033	.390	-.034

		$a(3)$	.297	-.023	.294	-.019
	.5	$a(1)$	.392	-.022	.395	-.030
		$a(2)$	.581	.010	.582	.012
		$a(3)$	.443	-.071	.444	-.064
	.8	$a(1)$	.620	-.010	.620	-.013
		$a(2)$	.653	.079	.652	.076
		$a(3)$	.711	.006	.710	.005
Matched	.2	$a(1)$	.299	-.006	.295	-.007
		$a(2)$	.213	-.015	.224	-.013
		$a(3)$	.165	-.026	.165	-.021
	.5	$a(1)$	.313	.009	.315	.012
		$a(2)$	.528	-.074	.523	-.073
		$a(3)$	.152	.044	.152	.042
	.8	$a(1)$	.510	.004	.507	.003
		$a(2)$	.312	-.029	.316	-.031
		$a(3)$	.268	.013	.266	.012
Hierarchical	.2	$a(1)$	.261	-.002	.261	-.001
		$a(2)$	.385	-.001	.383	-.007
		$a(3)$	.178	-.052	.179	-.056
	.5	$a(1)$	.193	-.032	.197	-.030
		$a(2)$	.216	-.120	.219	-.121

		$a(3)$	.230		.041	.230	.038
	.8	$a(1)$	.113	-	.103	.110	- .110
		$a(2)$	.132	-	.031	.132	- .025
		$a(3)$	.251	-	.194	.250	- .188

Table 4

RMSE and Bias of Intercept Estimates for the Four Transitional Points: γ(1), γ(2), γ(3), and γ(4) in the M2PPC Model When Test Length = 15

Prior	Inter-dimension Correlation	Parameters	Gibbs Sampler			HMC-NUTS		
			RMSE	Bias		RMSE	Bias	
Vague	.2	γ(1)	.240	-	.012	.240	-	.012
		γ(2)	.317	-	.027	.311	-	.028
		γ(3)	.209	-	.059	.210	-	.058
		γ(4)	.323	-	.028	.335	-	.028
	.5	γ(1)	.200	-	.022	.118	-	.022
		γ(2)	.432	-	.041	.430	-	.044
		γ(3)	.189	-	.027	.192	-	.025
		γ(4)	.297		.003	.303		.003
	.8	γ(1)	.452		.010	.455		.009

		γ(2)	.317	-.019	.308	-.019
		γ(3)	.303	-.032	.308	-.033
		γ(4)	.378	.008	.359	.007
Matched	.2	γ(1)	.179	-.001	.213	-.001
		γ(2)	.346	-.012	.372	-.011
		γ(3)	.128	-.013	.114	-.011
		γ(4)	.247	-.029	.228	-.030
	.5	γ(1)	.437	.004	.409	.022
		γ(2)	.301	.001	.300	.001
		γ(3)	.355	.021	.304	.017
		γ(4)	.197	-.009	.199	-.009
	.8	γ(1)	.131	-.004	.131	-.003
		γ(2)	.199	-.010	.198	-.008
		γ(3)	.277	-.015	.264	-.019
		γ(4)	.190	-.001	.191	-.001
Hierarchical	.2	γ(1)	.208	-.002	.201	-.001
		γ(2)	.214	-.001	.206	-.001
		γ(3)	.106	-.001	.112	-.001
		γ(4)	.121	-.042	.118	-.039
	.5	γ(1)	.310	-.003	.304	-.002
		γ(2)	.208	-.001	.211	-.001

		γ(3)	.179	- .001	.177	- .001
		γ(4)	.103	.001	.120	.001
	.8	γ(1)	.109	- .004	.109	- .002
		γ(2)	.189	- .001	.189	- .001
		γ(3)	.176	- .002	.174	- .001
		γ(4)	.103	- .008	.101	- .009

Table 5

RMSE and Bias of Slope Estimates on the Three Dimensions: a(1), a(2), and a(3) in the M2PPC Model When Test Length = 15

Prior	Inter-dimension Correlation	Parameters	Gibbs Sampler		HMC-NUTS	
			RMSE	Bias	RMSE	Bias
Vague	.2	$a(1)$	.204	- .100	.200	- .100
		$a(2)$	.321	- .026	.319	- .027
		$a(3)$	.282	- .017	.282	- .015
	.5	$a(1)$	.301	- .012	.300	- .011
		$a(2)$	.473	.010	.470	.010
		$a(3)$	.239	- .060	.233	- .052
	.8	$a(1)$	.468	- .008	.468	- .012

		$a(2)$	.589	.055	.587	.055
		$a(3)$	.492	.003	.489	.005
Matched	.2	$a(1)$	.198	-.004	.198	-.005
		$a(2)$	.175	-.011	.183	-.011
		$a(3)$	.132	-.018	.130	-.014
	.5	$a(1)$	.276	.007	.243	.007
		$a(2)$	.319	-.043	.317	-.044
		$a(3)$	.145	.027	.145	.027
	.8	$a(1)$	.322	.003	.319	.002
		$a(2)$	.199	-.017	.201	-.013
		$a(3)$	.187	.011	.189	.010
Hierarchical	.2	$a(1)$	.177	-.001	.174	-.001
		$a(2)$	.296	-.001	.287	-.001
		$a(3)$	.163	-.021	.163	-.026
	.5	$a(1)$	.145	-.020	.147	-.019
		$a(2)$	.199	-.080	.199	-.083
		$a(3)$	.221	.024	.220	.024
	.8	$a(1)$	.108	-.006	.100	-.006
		$a(2)$	.126	-.022	.130	-.020
		$a(3)$	.194	-.093	.194	-.087

Person Parameter Recovery

The RMSE and Bias values were averaged across all persons for evaluation of the recovery of person ability parameters on three related dimensions: $\theta(1)$, $\theta(2)$, and $\theta(3)$ in the M2PPC model using Gibbs Sampler and HMC-NUTS, and are summarized in Tables 6 and 7. The visual representations of person ability RMSEs: $\theta(1)$, $\theta(2)$, and $\theta(3)$ are summarized in Figures 11 through 13.

By inspection of Table 6 and Table 7, HMC-NUTS recovered the person parameters a little better than Gibbs Sampler with smaller *RMSE*s except for $\theta(2)$ when interdimensional correlation= .8 and test length=30 with Vague prior.

Similar to findings regarding test length in item parameter recovery, there was a consistent pattern of the *RMSE*s for θs. (From Figure 11 to Figure 13, the blue lines representing test length=30 are all below the orange lines representing test length=15.) A decrease in *RMSE*s was found with an increase in test length regardless of priors and interdimensional correlations, which means as the test length increases, the precision of person parameter recovery increases.

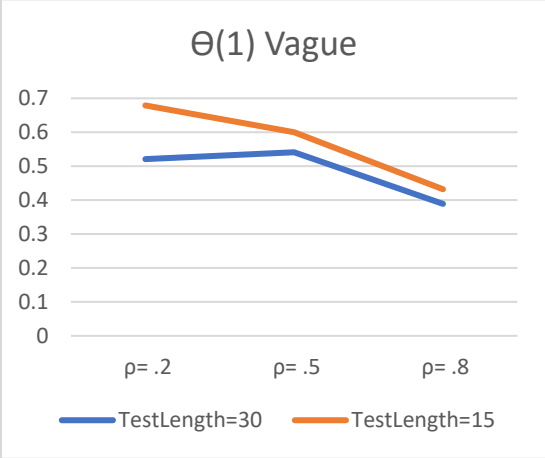
In terms of prior choice, Matched priors and Hierarchical priors recovered θs better than Vague priors. Additionally, Hierarchical priors recovered θs better than Matched priors with lower RMSEs for both algorithms and test lengths in all the three different interdimensional correlation conditions.

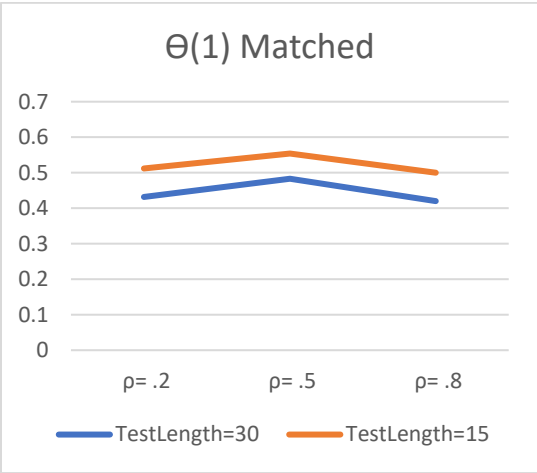
There was a trend for the influence of interdimensional correlation on θ recovery regardless of the test length, different algorithms and different prior choices. The precision of the person parameter recovery increased as the interdimensional correlation

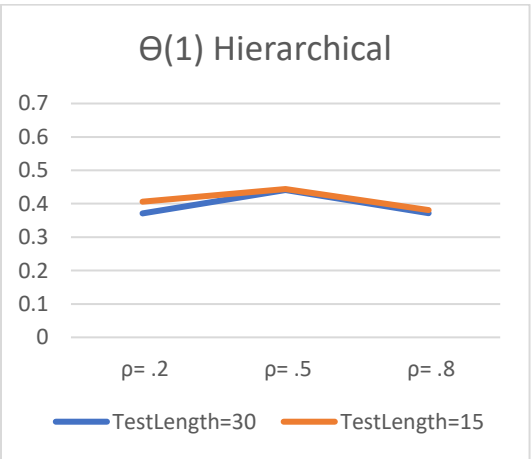
increased in all the different conditions for both algorithms (See Figure 11 through Figure 13).

The Bias estimates for person ability parameters were all positive except for the ones when interdimensional correlation = .2 using Matched priors. The majority of the person parameters were overestimated, which means the person with higher ability was estimated with an even higher ability. The three person parameters were underestimated when the interdimensional correlation = .2 using Matched priors, which means the person with higher ability was estimated with relatively lower ability.

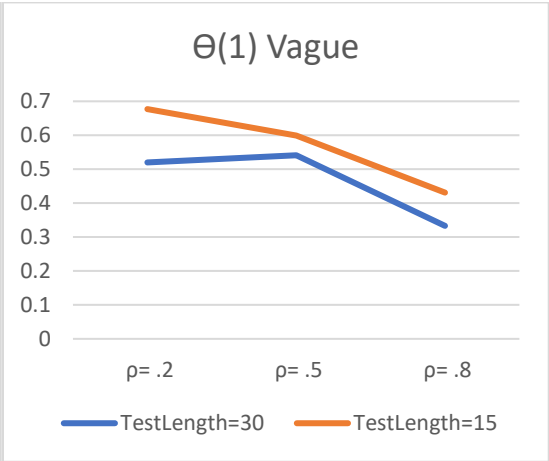
Gibbs Sampler

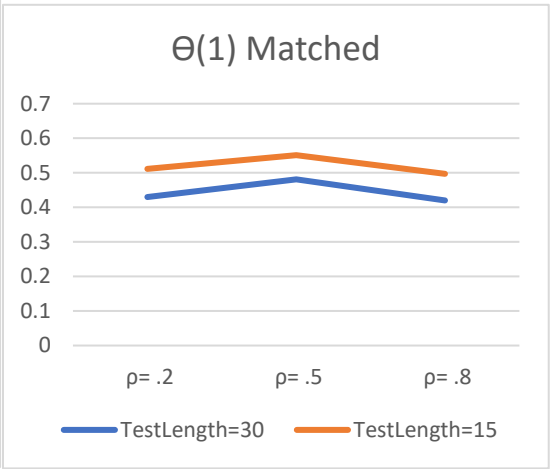






HMC-NUTS





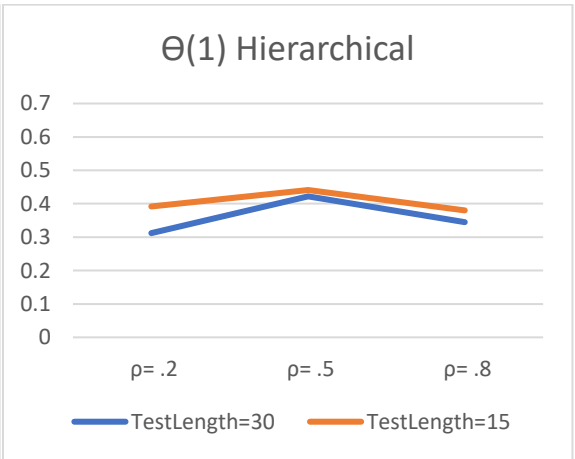
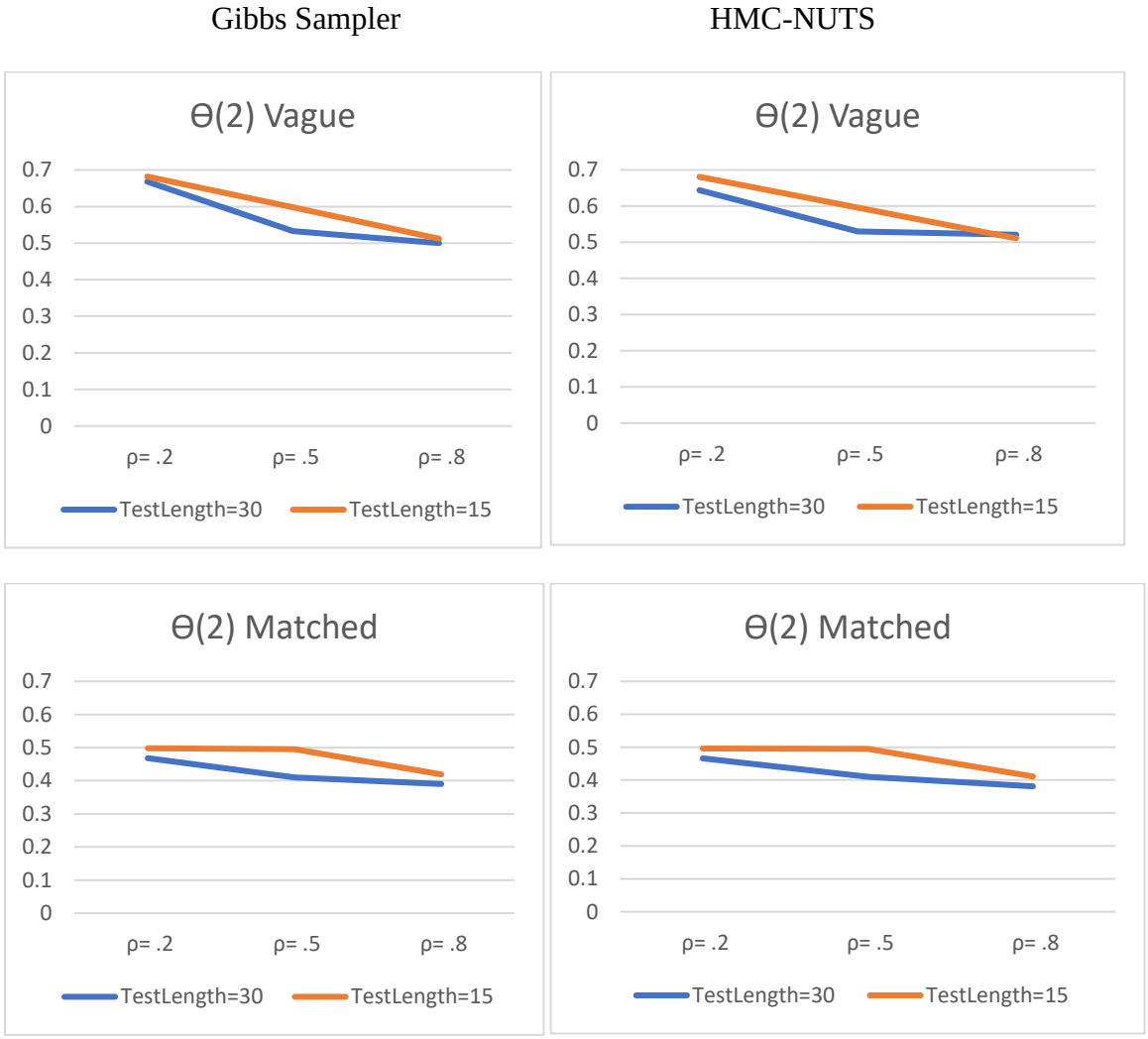
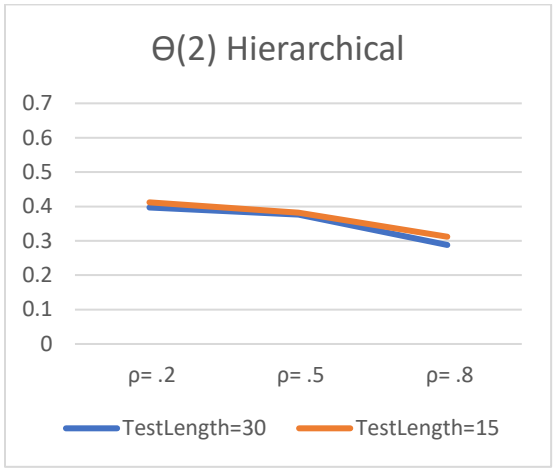


Figure 11. Average RMSE for Recovering Person Ability Parameter *θ(1)* Using Vague, Matched and Hierarchical Prior When Test Length Equal to 30 and 15 and Interdimensional Correlation Equal to .2, .5 and .8 in the M2PPC Model.





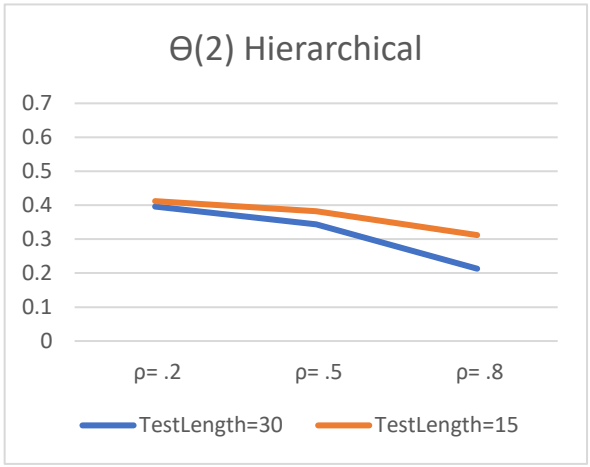
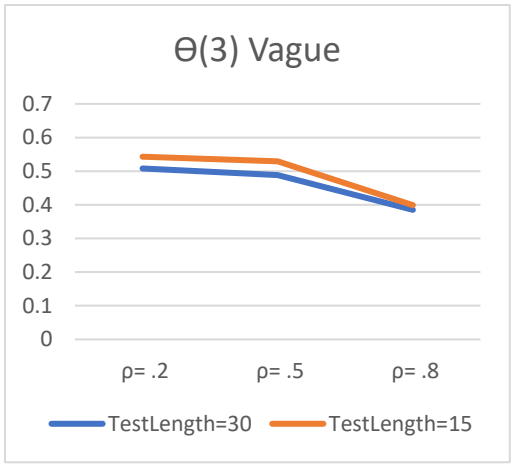
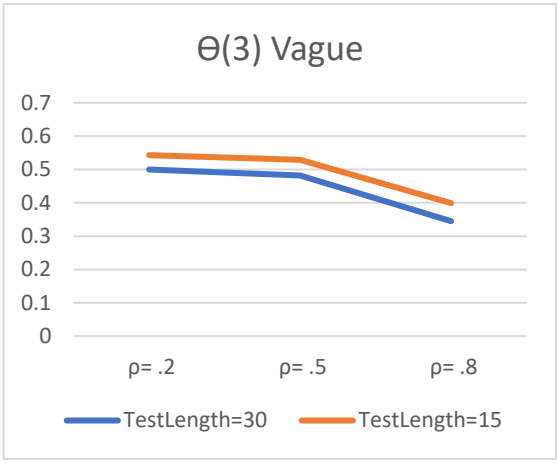


Figure 12. Average RMSE for recovering person ability parameter θ(2) using Vague, Matched and Hierarchical Prior when Test Length equal to 30 and 15 and interdimensional correlation equal to .2, .5 and .8 in the M2PPC model

Gibbs Sampler HMC-NUTS





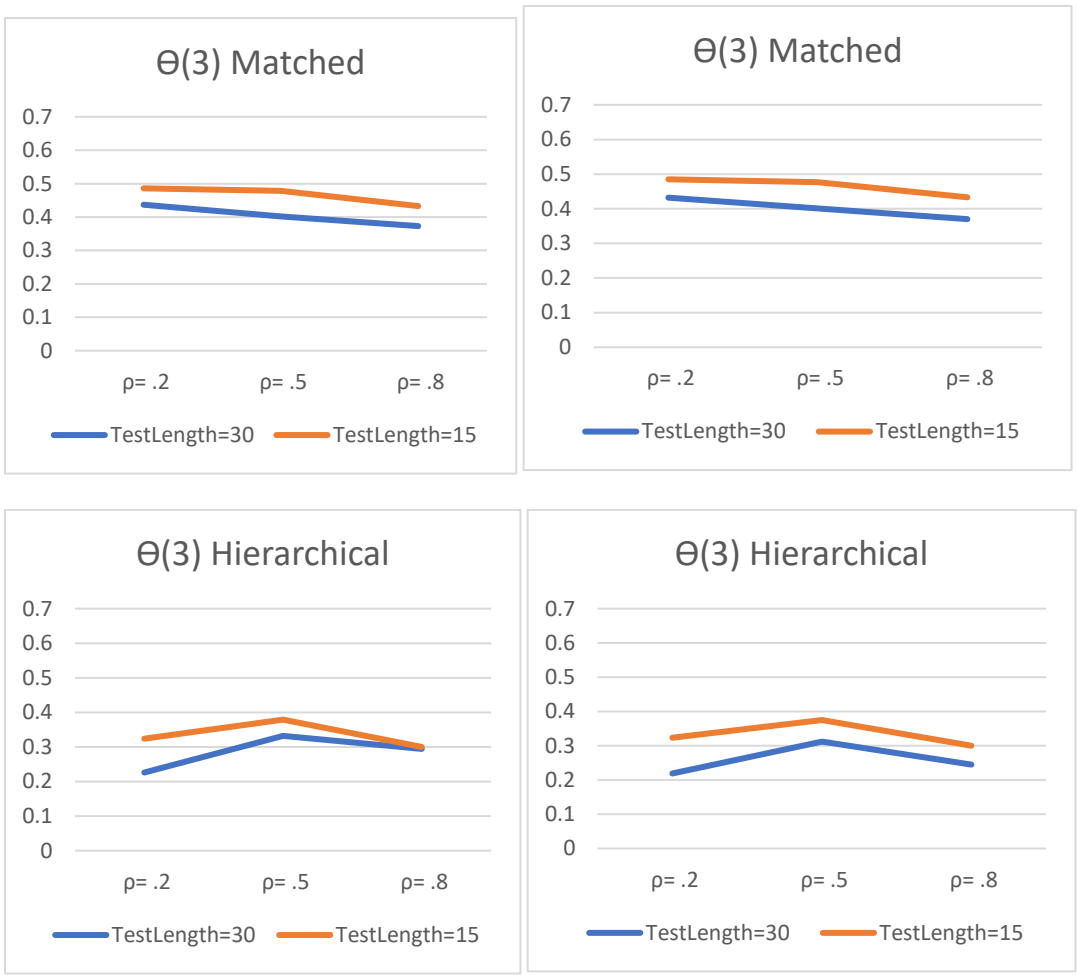


Figure 13. Average RMSE for recovering person ability parameter *θ(3)* using Vague, Matched and Hierarchical Prior when Test Length equal to 30 and 15 and interdimensional correlation equal to .2, .5 and .8 in the M2PPC model.

Table 6

RMSE and Bias of Person Ability Estimates on Three Dimensions: θ(1), θ(2), and θ(3) in the M2PPC Model When Test Length = 30

Prior	**Inter-dimension**	**Parameters**	**Gibbs Sampler**	**HMC-NUTS**

	Correlation					
			RMSE	Bias	RMSE	Bias
Vague	.2	$\theta(1)$	.521	.012	.520	.014
		$\theta(2)$	.668	.021	.644	.026
		$\theta(3)$	.508	.011	.500	.011
	.5	$\theta(1)$	.541	.012	.541	.010
		$\theta(2)$	.533	.013	.530	.011
		$\theta(3)$	.488	.035	.482	.032
	.8	$\theta(1)$	.389	.014	.333	.010
		$\theta(2)$	.500	.023	.521	.028
		$\theta(3)$	.385	.022	.345	.022
Matched	.2	$\theta(1)$	.432	-.003	.430	-.002
		$\theta(2)$	.468	-.001	.466	-.010
		$\theta(3)$	.437	-.003	.432	-.009
	.5	$\theta(1)$	.483	.001	.481	.001
		$\theta(2)$	.410	.002	.410	.001
		$\theta(3)$	.402	.007	.401	.003
	.8	$\theta(1)$	.420	.003	.420	.003
		$\theta(2)$	.390	.006	.381	.007
		$\theta(3)$	.373	.001	.370	.003
Hierarchial	.2	$\theta(1)$	.371	.001	.312	.002
		$\theta(2)$	.397	.001	.396	.001
		$\theta(3)$	.226	.001	.219	.001
	.5	$\theta(1)$	.441	.002	.422	.001
		$\theta(2)$	.376	.006	.344	.006
		$\theta(3)$	.332	.004	.312	.005

	.8	$\theta(1)$	.372	.006	.345	.007
		$\theta(2)$	.288	.002	.213	.003
		$\theta(3)$	.294	.003	.245	.001

Table 7

RMSE and Bias of Person Ability Estimates on Three Dimensions: θ(1), θ(2), and θ(3) in the M2PPC Model When Test Length = 15

Prior	**Inter-dimension Correlation**	**Parameters**	**Gibbs Sampler**		**HMC-NUTS**	
			RMSE	Bias	RMSE	Bias
Vague	.2	$\theta(1)$	.679	.023	.677	.022
		$\theta(2)$	.682	.031	.681	.032
		$\theta(3)$	.543	.020	.543	.022
	.5	$\theta(1)$	.600	.021	.599	.020
		$\theta(2)$	.598	.020	.596	.021
		$\theta(3)$	.529	.040	.529	.039
	.8	$\theta(1)$	.432	.020	.431	.018
		$\theta(2)$	.512	.031	.511	.031
		$\theta(3)$	.399	.030	.399	.030
Matched	.2	$\theta(1)$	.512	-.009	.511	-.009
		$\theta(2)$	.498	-.008	.496	-.010
		$\theta(3)$	.486	-.008	.485	-.006
	.5	$\theta(1)$	.554	.007	.551	.004
		$\theta(2)$	.495	.010	.495	.010

		$\theta(3)$	.479	.014	.477	.010
	.8	$\theta(1)$	.500	.011	.497	.011
		$\theta(2)$	.419	.012	.411	.012
		$\theta(3)$	.433	.011	.433	.011
Hierarchical	.2	$\theta(1)$	.406	.009	.392	.009
		$\theta(2)$	.412	.008	.412	.008
		$\theta(3)$	.324	.008	.323	.010
	.5	$\theta(1)$	.444	.009	.441	.009
		$\theta(2)$	.382	.014	.382	.013
		$\theta(3)$	.379	.011	.375	.010
	.8	$\theta(1)$	.381	.013	.380	.013
		$\theta(2)$	.312	.011	.312	.010
		$\theta(3)$	.300	.010	.300	.007

In summary, Gibbs Sampler and HMC-NUTS recovered the item parameters similarly for all simulated conditions, but HMC-NUTS recovered the person parameter better than Gibbs Sampler with only one exception (which was $\theta(2)$ when interdimensional correlation= .8 and test length=30 with Vague prior). As the test length increased the precision decreased, but for person parameters, the impact of test length was in the opposite direction--as the test length increased the precision increased. In addition, the higher the interdimensional correlation, the more precisely the person parameters were recovered in all simulated conditions. Matched priors and Hierarchical priors both recovered the item parameters more precisely than Vague priors, and the Hierarchical priors recovered the person parameter most precisely among the three priors. Both algorithms show positive Bias values for person parameter recovery except for one

condition (interdimensional correlation = .2 using Matched priors), which means most of the persons with higher ability were estimated to be even higher.

Computational Speed of Gibbs Sampler and HMC-NUTS

Regarding the computational speed for implementing Gibbs Sampler and HMC-NUTS respectively in *rjags* and *rstan*, two methods were utilized: a personal computer (PC) with a processor 2.7 GHz with Turbo Boost Intel Core i7 and memory 16 GB 1866 MHz and a High Performance Computer (HPC) cluster from the University of Denver (http://portfolio.du.edu/du_hpc/page/47530) with 44 compute nodes and 456 available computational cores, which can be utilized for parallel computing for different chains .

Since the M2PPC model is a complicated model, the PC was only used for estimating one of the simulated datasets, where the condition was: Matched prior, test length=30, and interdimensional correlation= .2. Gibbs sampler via *rjags* took 359 minutes to complete the three chains with 11,000 iterations, but HMC-NUTS via *rstsan* took 562 minutes to complete the same task.

When it came to utilizing the HPC, the computational speed varied for different conditions. Overall, the shorter the test length, the faster the computational speed. Matched priors and Hierarchical priors took similar computational times and both were faster than Vague priors in all the simulated conditions. For example, when the condition was the same in using the PC-Matched prior, test length=30, and interdimensional correlation= .2, Gibbs Sampler via *rjags* took 181 minutes to complete the three chains with 11,000 iterations, but HMC-NUTS via *rstan* took 309 minutes to complete the same task. Moreover, implementing HMC-NUTS via *rstan* code had some unclear mixing

issues in some of the conditions when using Vague priors, which needs further investigation and modification to make it more time efficient.

Analysis of Variance (ANOVA) Results

Log*RMSE*

Effect sizes (ω^2) of log*RMSE* for the four factors, namely, test length, interdimensional correlations, priors, and two MCMC algorithms in estimating parameters of M2PPC model are summarized in Table 8 and Table 9.

For intercept parameter estimation, Table 8 shows that test length had the largest effect for all the γ(1), γ(2), γ(3), and γ(4) estimates compared to the other factors, explaining about 19.4%, 20.5%, 12.7% and 16.6% respectively, of the total variance in log*RMSE*. The effect sizes for γ(1), γ(2), and γ(4) were large and the effect size for γ(3) was medium, according to Cohen (1988). Prior choice had the second to the largest effect on the estimation of γ(1), γ(2), γ(3), and γ(4), accounting for 14.2%, 15.1%, 10.9% and 13.2% of the total variance in log*RMSE*. The effect sizes for γ(1) and γ(2) were large, but the effect sizes for γ(3) and γ(4) were medium based on Cohen (1988). All the other main effects and interaction contributed less than or equal to 1% of the total variance in log*RMSE*.

For slope parameter estimation, Table 8 shows that test length also had the largest effect for all the $a(1)$, $a(2)$, and $a(3)$ estimates compared to the other factors, explaining about 21.3%, 18.9%, and 15.9% respectively, of the total variance in log*RMSE*. The effect sizes for $a(1)$, $a(2)$, and $a(3)$ were all large according to Cohen (1988). Prior choice had the second to the largest effect on the estimation of $a(1)$, $a(2)$, and $a(3)$,

accounting for 10.4%, 16.0%, and 17.4% of the total variance in log*RMSE*. The effect sizes for $a(2)$ and $a(3)$ were large, but the effect size for $a(1)$ was medium based on Cohen (1988). All the other main effects and interaction contributed less than or equal to 1% of the total variance in log*RMSE*.

Table 8

Effect Sizes (ω^2) for Main Effects and Interactions on Log*RMSE of Item Parameter Estimates in the M2PPC Model*

Source of Variation	**γ(1)**	**γ(2)**	**γ(3)**	**γ(4)**	$\boldsymbol{a}$**(1)**	$\boldsymbol{a}$**(2)**	$\boldsymbol{a}$**(3)**
Test Length	.194	.205	.127	.166	.213	.189	.159
Correlation	.001	.004	.000	.001	.001	.003	.000
Prior	.142	.151	.109	.132	.104	.160	.174
Algorithm	.000	.001	.000	.000	.000	.000	.000
Test Length*Correlation	.002	.001	.000	.003	.002	.006	.002
Test Length*Prior	.008	.010	.009	.010	.004	.011	.009
Test Length*Algorithm	.005	.004	.000	.001	.004	.005	.003
Correlation*Prior	.003	.000	.001	.003	.002	.001	.001
Correlation*Algorithm	.000	.000	.000	.000	.000	.000	.000
Prior*Algorithm	.004	.000	.000	.001	.007	.003	.001
Test Length*Correlation*Prior	.001	.000	.000	.000	.001	.000	.000
Test Length*Correlation*Algorithm	.000	.000	.000	.000	.000	.000	.000
Test Length*Prior*Algorithm	.001	.000	.000	.000	.000	.001	.001
Correlation*Prior* Algorithm	.000	.000	.000	.000	.000	.000	.000

Test Length*Correlation*Prior*Algorithm	.000	.000	.000	.000	.000	.000	.000

In terms of person parameter estimates, Table 9 shows that test length again explained the largest amount of variance in log*RMSE*, namely 26.1%, 27.4% and 34.1% for θ(1), θ(2) and θ(3). Prior choice also contributed 14.9%, 15.1% and 14.3% of the total variance in log*RMSE* for θ(1), θ(2) and θ(3). Tength and prior choice both had large effects on the estimation of person parameters. Furthermore, different from that in item parameter estimation, interdimensional correlation explaineda medium amount of variance in log*RMSE*, 11.2%, 10.8% and 11.7% respectively for θ(1), θ(2) and θ(3). The two different MCMC algorithms accounted for small amounts of variance in log*RMSE*, 4.3%, 3.7% and 4.0% for θ(1), θ(2) and θ(3). Apart from these main effects, the interaction between interdimensional correlation and prior choice also explained small amounts of variance in the total variance of log*RMSE*, 1.4%, 1.5% and 1.1% for θ(1), θ(2) and θ(3). All the other interactions had almost no effect on log*RMSE*.

Table 9

Effect Sizes (ω^2) for Main Effects and Interactions on Log*RMSE of Person Parameter Estimates in the M2PPC Model*

Source of Variation	θ(1)	θ(2)	θ(3)
Test Length	.261	.274	.341
Correlation	.112	.108	.117
Prior	.149	.151	.143

Algorithm	.043	.037	.040
Test Length*Correlation	.009	.010	.004
Test Length*Prior	.002	.007	.008
Test Length*Algorithm	.004	.003	.009
Correlation*Prior	.014	.015	.011
Correlation*Algorithm	.001	.000	.001
Prior*Algorithm	.009	.008	.005
Test Length*Correlation*Prior	.002	.000	.001
Test Length*Correlation*Algorithm	.000	.000	.001
Test Length*Prior*Algorithm	.000	.000	.000
Correlation*Prior* Algorithm	.000	.000	.000
Test Length*Correlation*Prior*Algorithm	.000	.000	.000

Log*Bias*

Effect sizes (ω^2) of log*Bias* for the four factors, namely, test length, interdimensional correlation, priors, and two MCMC algorithms in estimating parameters of M2PPC model are summarized in Table 10 and Table 11.

For item parameter estimation, test length had small effects on log*Bias* for γ(1), γ(2), and γ(4), accounting for 2.2%, 1.5% and 1.3% of the total variance, and small effects on log*Bias* for a(1) and a(2), accounting for 2.1% and 3.3% of the total variance. Interdimensional correlation also had small effects on log*Bias* for γ(2) and γ(3), accounting for 1.2% and 3.0% of the total variance, and small effects on log*Bias* for a(1),

accounting for 1.4% of the total variance. Lastly, Prior Choice had small effects on log*Bias* for γ(1), γ(2), and γ(3), accounting for 1.5%, 2.2% and 1.0% of the total variance, and small effects on log*Bias* for a(1) and a(2), accounting for 1.2% and 1.8% of the total variance. All the other main effects and interactions contributed less than 1% of the total variance of log*Bias*.

Table 10

Effect Sizes (ω^2) for Main Effects and Interactions on Log*Bias of Item Parameter Estimates in the M2PPC Model*

Source of Variation	γ(1)	γ(2)	γ(3)	γ(4)	a(1)	a(2)	a(3)
Test Length	.022	.015	.009	.013	.021	.033	.005
Correlation	.000	.012	.030	.006	.014	.000	.001
Prior	.015	.022	.010	.009	.012	.018	.004
Algorithm	.000	.001	.000	.000	.000	.000	.000
Test Length*Correlation	.001	.001	.000	.003	.002	.006	.000
Test Length*Prior	.001	.000	.000	.000	.001	.000	.001
Test Length*Algorithm	.000	.000	.000	.000	.000	.001	.000
Correlation*Prior	.000	.000	.000	.000	.000	.001	.001
Correlation*Algorithm	.000	.000	.000	.000	.000	.000	.000
Prior*Algorithm	.004	.000	.000	.001	.001	.000	.001
Test Length*Correlation*Prior	.000	.000	.000	.000	.000	.000	.000
Test Length*Correlation*Algorithm	.000	.000	.000	.000	.000	.000	.000
Test Length*Prior*Algorithm	.001	.000	.000	.000	.000	.000	.000

Correlation*Prior* Algorithm	.000	.000	.000	.000	.000	.000	.000
Test Length*Correlation*Prior*Algorithm	.000	.000	.000	.000	.000	.000	.000

Regarding person parameter estimation, interdimensional correlation had small effects on log*Bias* for θ(2) and θ(3), accounting for 1.6% and 1.9% of the total variance. Prior choice had small effects on log*Bias* for θ(1), θ(2), and θ(3), accounting for 2.6%, 1.1% and 1.0% of the total variance. All the other main effects and interactions contributed less than 1% of the total variance of log*Bias*.

Table 11

Effect Sizes (ω^2) for Main Effects and Interactions on Log*Bias of Person Parameter Estimates in the M2PPC Model*

Source of Variation	θ(1)	θ(2)	θ(3)
TestLength	.001	.006	.000
Correlation	.005	.016	.019
Prior	.026	.011	.010
Algorithm	.001	.001	.001
TestLength*Correlation	.000	.000	.000
TestLength*Prior	.000	.000	.000
TestLength*Algorithm	.000	.000	.000
Correlation*Prior	.001	.000	.000

Correlation*Algorithm	.001	.000	.001
Prior*Algorithm	.000	.000	.000
TestLength*Correlation*Prior	.000	.000	.000
TestLength*Correlation*Algorithm	.000	.000	.000
TestLength*Prior*Algorithm	.001	.000	.000
Correlation*Prior* Algorithm	.000	.000	.000
TestLength*Correlation*Prior*Algorithm	.000	.000	.000

In summary, the ANOVA results supports the conclusions that can be drawn from Table 2 through Table 7. Test length plays an influential role in estimating both item and person parameter estimation in log*RMSE*, since it explains the majority of the total variance in log*RMSE*. Prior choice also affects both item and person parameter estimation in log*RMSE* based on the variance explained by the three different prior choices. In terms of person parameter recovery, the interdimensional correlation also contributes to the mount of varianceexplained in log*RMSE*, which is in line with the conclusion drawn previously, namely, the precision of the person parameter estimation increases as the interdimensional correlation increases. Regarding log*Bias*, interdimensional correlation and prior choice both showed small effects on person parameter estimation in log*Bias*, which is consistent with the conclusion that the Bias values of all the person parameter estimates are positive except for one situation with interdimensional correlation = .2 using Matched prior. However, test length,

interdimensional correlation, and prior choice all have small effects on item parameter estimation in log*Bias*, which is not explcitly shown based on Tables 2 through 5.

Chapter Four: Findings and Discussion

Introduction

This chapter includes four main sections. The first section summarizes the results of the application of Gibbs Sampler and Hamiltonian Monte Carlo-No-U-Turn Sampler (HMC-NUTS) for estimating parameters in the Multidimensional Two Parameter Partial Credit Model (M2PPC). Also, the answers to the three research questions are synthesized in this section. Implications are discussed for using the fully Bayesian estimation methods with different test lengths, prior choices, and interdimensional correlations. Then, the second section provides a discussion of how study results related to the existing literature. Finally, the limitations and directions for future studies are presented.

The Findings of This Study

This simulation study compared Gibbs Sampler and HMC-NUTS bias and RMSE in the parameters for the M2PPC model via manipulating three factors: test length, prior choice, and interdimensional correlation. When considering the computational speed and estimation accuracy, the results of parameter recovery of the M2PPC model show that Gibbs Sampler and HMC-NUTS performed similarly in all the simulated conditions. Based on the conclusions from Chapter 3, for item parameter estimation, Gibbs Samplers and HMC-NUTs did differed only slightly in their RMSE and Bias; for person parameter recovery, HMC-NUTS performed a little better than

Gibbs Sampler with smaller RMSEs. However, considering the computational speed, the Gibbs Sampler recovered the M2PPC model parameters significantly faster than HMC-NUTS. Moreover, HMC-NUTS implemented via *rstan* had some unclear chain mixing issues for some iterations. The significant differences in computing time for these two algorithms may be due to implementing them via two different packages (*rjags* and *rstan*). It would be more appropriate to compare them via the same computational package, which is difficult based on the current development of computer hardware. Practitioners might take the estimation precision and computational speed into consideration when choosing one of the two estimation methods.

Concerning test length, there was a consistent pattern for RMSEs that the accuracy of item parameter estimates increased as the test length decreased, but the accuracy of person parameter estimates increased as the test length increased in all the simulated conditions for both Gibbs Sampler and HMC-NUTS. As test length decreases, the number of items that need to be recovered drops when the sample size remains the same. So with shorter tests, each item parameter is recovered more accurately. As the test length increases, there is more information coming from the items collected to predict the person abilities; thus the accuracy of person ability estimates increases. Test length had no consistent impact on Bias for either item parameter or person parameter estimates.

Different interdimensional correlations did not influence the recovery of item parameters but affected the precision of the estimation of person parameters. The accuracy of the person parameter recovery increased as the interdimensional correlation increased in all the different conditions for both Gibbs Sampler and HMC-NUTS.

Increased interdimensional correlation indicates the person's latent traits are sharing more information with one another; therefore, when other features remain the same, the information collected by the same number of items will increase, which ultimately will improve the precision of person parameter estimates. For instance, if a scale with ten items measures depression and anxiety with sample size equal to 100, the accuracy of person parameter estimates will be more accurate than a scale with ten items measuring depression and sleeping problems with the same sample size, since the correlation between depression and anxiety is assumed to be higher than the correlation between depression and sleeping problems.

Gibbs Sampler and HMC-NUTS with standard Vague priors yielded the least accurate estimates for both item and person parameters. More specifically, Matched priors and Hierarchical priors recovered item parameters and person parameters similarly and both were better than Vague priors. Furthermore, Hierarchical priors recovered person parameters the best among the three different priors. Even though Matched priors results were very similar to Hierarchical priors for item parameter recovery, Matched priors is generally unavailable in real estimation applications since the actual distribution of parameters is unknown. Matched priors would be useful in situations where previous research can provide reliable evidence about the parameter distributions. Practitioners may want to use Hierarchical priors no matter which one of the two estimation methods is chosen as it is more time-efficient and as precise. Lastly, both Gibbs Sampler and HMC-NUTS recovered the slope parameters better, with smaller RMSEs, than the intercept parameters in all conditions.

The results of analyses of variance (ANOVAs) supported the conclusions drawn previously. Test length and prior choice both accounted for a large amount of total variance in log*RMSE* for both item and person parameter recovery. Also, interdimensional correlation explained a medium amount of variance in log*RMSE* for person parameters. And, interdimensional correlation and prior choice accounted for a small amount of variance in log*Bias*.

Discussion

In general, this study examined the performance of two full Bayesian estimation methods: Gibbs Sampler and HMC-NUTS in estimating a M2PPC model under different conditions using simulated data. Only one finding of this study contradicts the existing literature. According to Martin-Fernandez and Revuelta (2017), HMC-NUTS converges faster than Gibbs Sampler with sample size equal to 500 and test length equal to 18 or 25 items. However, in the present study, Gibbs Sampler converged more quickly than HMC-NUTS with sample size equal to 500 and test length equal to 15 or 30 items. This contradiction may be due to the different levels of IRT model complications. Martin-Fernandez and Revuelta (2017) used a dichotomous multidimensional IRT model, but the current study used a polytomous multidimensional IRT model.

Addressing the gap stated in Martin-Fernandez and Revuelta's (2017) simulation study that more precise estimates could be obtained when sample size and test length for each dimension increased, without clearly talking about the item and person parameter respectively, this study showed a consistent pattern for RMSEs in that the accuracy of item parameter estimates increased as the test length decreased, but the accuracy of

person parameter estimates increased as the test length increased in all the simulated conditions for both Gibbs Sampler and HMC-NUTS.

This study extends the findings from Natesan et al. (2016) that in unidimensional 1-PL and 2-PL dichotomous models, the hierarchical priors and matched priors performed better than vague priors, and also the vague priors produced large errors or convergence issues in parameter recovery using Gibbs Sampler and Variational Bayesian estimation methods--which are not recommended. Similarly, the current study found that in estimating the multidimensional polytomous 2-PL model, the hierarchical priors and matched priors still performed better than vague priors.

The current study also advances conclusions regarding the influence of interdimensional correlation on IRT model parameter estimation. In a 2-PL graded response model, Kuo and Sheng (2016) found that when the interdimensional correlation (only including two dimensions) was low, the estimating methods (including Marginal Likelihood estimation methods and Fully Bayesian Estimation methods) provided similar results. The current simulation study demonstrates that as the interdimensional correlation increases, the accuracy of person parameter estimation will also increase when there are three correlated dimensions. The findings concerning the impact of interdimensional correlation on parameter estimation in a M2PPC model addresses a gap in the literature and also poses new research questions about higher dimensional IRT models. For instance, with higher dimensional IRT models, such as four and five related dimensions, is this trend still true?

Limitations

Through simulation studies, this dissertation demonstrates that researchers and practitioners would benefit from using Gibbs Sampler with Hierarchical priors to estimate the parameters in M2PPC model, which is accessible, fast, and accurate.

It is, however, noted that all the conclusions are based on simulated conditions, which cannot necessarily be generalized to other situations. The current study used one sample size (n = 500), two test lengths: 15 and 30 items, three interdimensional correlations: .2, .5, and .8, a five-point Likert scale, uncorrelated discrimination parameters (slope parameters) on the three dimensions, and an equal number of items for three related dimensions in the M2PPC model. And the results of this study were based on only ten replications for all the different simulated conditions. Considering the small number of replications, the RMSE and Bias values presented in Chapter 3 need further verification with more studies before generalizing the results to other similar conditions.

One of the limitations is using the inverse-Wishart distribution as the prior for estimating the person parameter covariance matrix. Some studies, such as Alvarez, Niemi, and Simpson (2014) showed that an inverse-Wishart distribution might not work well as the prior in some situations. Specifically the prior does not work well when the actual variance is small compared to the prior mean. In this situation the posterior variance for the person ability parameter estimates will be biased. Another concern when using inverse-Wishart as the prior is that it impacts the estimation accuracy of all the parameters' variances by setting a single degree of freedom, so that the marginal distribution of the variances have densities near zero (Gelman, 2014). In this current

study, almost all the person abilities parameters were overestimated, which likely is a consequence of using the inverse-Wishart prior. Because of these reasons, the Stan manual (Stan Development Team, 2016) suggests LKJ Cholesky Covariance priors for the covariance matrix (For more information about LKJ prior, please refer to the Stan Mannual. LKJ priors are based on work by Daniel Lewandowski, Dorota Kurowicka, and Harry Joe, 2009.) However, LKJ priors cannot be implemented in Gibbs Sampler via *rjags*, which is the major disadvantage.

Another limitation of this study is the lack of discussion of effective sample size (ESS) of MCMC chains. Most of the MCMC chains are highly autocorrelated, which means that the successive steps are not independent but are strongly correlated with each other. ESS is a measure of chain length taking the autocorrelation of the chain into consideration. The decision about how large the ESS should be for a study is heuristic, and depends on which details of the posterior distribution are of concern. In this study, the univariate ESSs fell in a range from 600 to 2,400. Since this was a simulation study, the ESSs were not inspected. Careful examination of ESSs is needed when estimating the M2PPC model in applications with real data.

Directions for Future Studies

The recently developed Metropolis-Hasting Robbins-Monro (MHRM; Cai, 2010a, 2010b) algorithm, which combines fully Bayesian estimation with a Robins-Monro technique to facilitate the maximum likelihood estimation, has shown advantages over traditional estimation methods, such as Gibbs Sampler, in estimating multidimensional

IRT models (Martin-Fernandez & Revuelta, 2017). Future studies may consider including this algorithm in comparison studies.

Simulation studies are often performed in manipulated and so ideal situations. In the current study, the sample size was fixed to 500 for computational convenience. Future studies may include a smaller sample size to explore the performance of the two Bayesian algorithms in low sample size situations. Also, the test lengths in this study were fixed at 15 and 30, and the estimations were conducted separately. It would be interesting to investigate: first splitting the 30-item test into two 15-item tests cases or three 10-item tests, then aggregrating the results, and finaly comparing the results with the 30-item test. By doing so, we could determine if the result of decreased item RMSE with shorter tests held, and so RMSE was lower with the aggregated result. Moreover, the current study only used three fixed dimensions, and future studies may incorporate more dimensions as the computational hardware continues to develop. In this specific case, the IRT M2PPC model was defined as known, and fit could be assumed to be almost perfect, which is not the case in real data applications. Future studies may use the two Bayesian estimation methods to fit M2PPC models and for further model comparison and selection in real data situations. There are a large number of prior choices that remain unexplored, such as scaled inverse-Wishart and LKJ priors, and they need to be explored for the person ability parameter estimates when using Hierarchical priors. Meanwhile, there are also numerous choices of simulated values for the IRT model parameters; future studies may use these two Bayesian estimation methods to fit IRT models with non-normal latent trait distributions, such as a gamma distributions. Lastly, for the model evaluation criteria,

other evaluation metrics can be incorporated in future studies as well, such as area under the curve (AUC).

References

Alvarez, I., Niemi, J., & Simpson, M. (2016). Bayesian inference for a covariance matrix. *ArXiv.org*, Jul 8, 2016.

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika*, *65*(4), 475-496.

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

Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., & West, M. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, 7, 453-464.

Bishop, C. (2006). *Pattern recognition and machine learning.* (Information science and statistics). New York: Springer.

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

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

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33-57.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307-335.

Camilli, G. (1994). Teacher's corner: origin of the scaling constant d=1.7 in item response theory. *Journal of Educational Statistics*, *19*(3), 293-295.

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

Fox, J. (2010). *Bayesian item response modeling : Theory and applications.* New York, NY: Springer.

Fu, Z., Tao, J., & Shi, N. (2010). Bayesian estimation of the multidimensional graded response model with nonignorable missing data. *Journal of Statistical Computation and Simulation*, 80(11), 1237-1252.

Gelman, A. (2014). *Bayesian data analysis* (3rd Ed.). Boca Raton, FL: CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.

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

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.

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

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*(1), 1593-1623.

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, *7*, 109.

Jones & X. L. Meng (Eds.), *Handbook of Markov chain Monte Carlo*. Boca Raton, FL:CRC Press.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79-86.

Kuo, T. C., & Sheng, Y. (2016). A comparison of estimation methods for a multi-unidimensional graded response IRT model. *Frontiers in Psychology*, *7*, 880.

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. New York: Chapman and Hall/CRC.

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

Martin-Fernandez, M., & Revuelta, J. (2017). Bayesian estimation of multidimensional item response models. A comparison of analytic and simulation algorithms. *Psicológica*, *38*(1).

Muraki, E. (1997). A generalized partial credit model. *Handbook of modern item response theory*, 153-164.

Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. D. (2016). Bayesian prior choice in IRT estimation using MCMC and variational Bayes. *Frontiers in Psychology*, *7*, 1422. http://doi.org/10.3389/fpsyg.2016.01422

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*(2), 287-312.

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research*, *43*(1), 73-97.

R. Philip Chalmers (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1-29.<doi:10.18637/jss.v048.i06>

Stan Development Team. (2018). *Stan Modeling Language Users Guide and Reference Manual*, Version 2.18.0. http://mc-stan.org

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*(393), 82--86.

Tsutakawa, R. K., & Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika*, *51*(2), 251-267.

Yanyan Sheng, C., & Wikle. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, *68*(3), 413-430.

Yanyan Sheng, & Todd C. Headrick. (2012). A Gibbs Sampler for the Multidimensional Item Response Model. *ISRN Applied Mathematics*, 2012(2012), 1-14.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*(3), 175-191.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236-247.

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

Wollack, J., Bolt, D., Cohen, A., & Young-Sun, L. (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.

Appendix A

R*jags* Script for Estimation in the M2PPC Model

Matched Prior, Interdimensional Correlation = .8, Test length=30

```
 # Assemble data into list for JAGS:
InvSigmaSubjAbil<-solve(matrix(rep(.8,9),nrow=3)+diag(.2,3))
zero3<-rep(0,3)
y = theData[,"Ans"]
itemID =rep(1:30, 500)
subjID = theData[,"Subj"]
Nitem = length(unique(itemID))
Nsubj = length(unique(subjID))
NAnsK = length(unique(y))
NDim = nrow(InvSigmaSubjAbil)
Ntotal = nrow(theData)
dataList = list(
  y=y , itemID=itemID , subjID=subjID , Nitem=Nitem , Nsubj=Nsubj ,
NAnsK=NAnsK,NDim=NDim,
  Ntotal=Ntotal, InvSigmaSubjAbil=InvSigmaSubjAbil,zero3=zero3)
#Speicify different initial values for the chains
initsList <-list()
set.seed(123456)
for (i in 1:3){
```

```
initsAbil = matrix(rnorm(500*3),nrow=500)
initsDisc = matrix(exp(rnorm(30*3)),nrow=30)
initsDiff = matrix(rnorm(30*4),nrow=30)
thisList=list(subjAbil=initsAbil,itemDiff=initsDiff,itemDisc=initsDisc,.RNG.name=
"base::Super-Duper",
                .RNG.seed=123456+length(initsList))
initsList[[length(initsList)+1]] <- thisList
}
## Define the model (Matched Prior):
modelString ="
model{
for ( rowIdx in 1:Ntotal ) {
  y[rowIdx] ~ dcat(pAns[rowIdx,1:NAnsK] )
}
# pAns has the probabilities of the answer categories for eqn. 1.5
for ( rowIdx in 1:Ntotal ) {
  for (AnsKx in 1:NAnsK){
     pAns[rowIdx,AnsKx]<-pMat[rowIdx,AnsKx]/sum(pMat[rowIdx,1:NAnsK])
    }
  }
for ( rowIdx in 1:Ntotal ) {
  pMat[rowIdx,1]<-1
```

```
  for (AnsKx in 2:NAnsK){
     pMat[rowIdx,AnsKx]<- pMat[rowIdx,AnsKx-1]*
        exp(inprod(itemDisc[itemID[rowIdx],1:NDim],subjAbil[subjID[rowIdx],1:NDim])+
                          itemDiff[itemID[rowIdx],(AnsKx-1)])
    }
}
  for ( subjIdx in 1:Nsubj ) {
    subjAbil[subjIdx,1:NDim] ~ dmnorm(zero3,InvSigmaSubjAbil)
  }
  for ( itemIdx in 1: Nitem ) {
    for (AnsKx in 1:(NAnsK-1)){
    itemDiff[itemIdx,AnsKx] ~ dnorm( 0 , 1 )
    }
  }
  for ( itemIdx in 1:Nitem ) {
    for (NDimx in 1:NDim){
    itemDisc[itemIdx,NDimx] ~ dlnorm( 0, 1/0.25 )
    }
  }
}
"
writeLines( modelString , con="532_1M.txt" )
```

```
# Run the chains:
library(runjags)
library(rjags)


runJagsOutM <- run.jags(
                    model="532_1M.txt" ,
                    monitor=c("subjAbil","itemDiff","itemDisc"),
                    data=dataList ,
                    inits=initsList ,
                    n.chains=3 ,
                    adapt=1000 ,
                    burnin=5000 ,
                    sample=ceiling(15000/3) ,
                    thin=1,
                    summarise=FALSE ,
                    plots=FALSE )
codaSamplesM = as.mcmc.list(runJagsOutM )
save(codaSamplesM , file="codaSamplesM.Rdata")
#write the estimates to csv
parameter_names <- varnames(codaSamplesM)
saved_steps <- as.integer(row.names(codaSamplesM[[1]]))
```

```
outM <- data.frame("chain" = factor(rep(1 : length(codaSamplesM), each =
length(saved_steps))),
                   "step" = rep(saved_steps, length(codaSamplesM)) )
outM <- cbind(outM, as.data.frame(as.matrix(codaSamplesM)))
write.csv(outM, "outM.csv")
```

Appendix B

Rstan Script for Estimation in the M2PPC Model

Hierarchical Prior, Interdimensional Correlation = .2, Test length=30

```
#SigmaSubjAbil<- matrix(rep(.2,9),nrow=3)+diag(.8,3)
zero3<-rep(0,3)
y = theData[,"Ans"]
itemID =rep(1:30, 500)
subjID = theData[,"Subj"]
Nitem = length(unique(itemID))
Nsubj = length(unique(subjID))
NAnsK = length(unique(y))
NDim = 3
Ntotal = nrow(theData)
W = diag(3)

dataList = list(
  y=y , itemID=itemID , subjID=subjID , Nitem=Nitem , Nsubj=Nsubj ,
NAnsK=NAnsK,NDim=NDim,
  Ntotal=Ntotal,zero3=zero3,W=W)
modelString = "
data {
int<lower=1> Nsubj;               // number of students
```

```
int<lower=1> Nitem;              // number of questions
int<lower=1> Ntotal;             // number of observations
int<lower=1,upper=Nsubj> subjID[Ntotal];    // student for observation n
int<lower=1,upper=Nitem> itemID[Ntotal];    // question for observation n
int<lower=1,upper=5> NAnsK;          //number of answer categories
int<lower=1,upper=5> y[Ntotal];     // category of observation n  y[N]
int<lower=1,upper=3> NDim;               // number of latent dimensions   D
vector[3] zero3;
matrix[3,3] W;
}
parameters {
matrix[Nitem,NAnsK-1] itemDiff;       //intercept parameters
matrix<lower=0>[Nitem,NDim] itemDisc;       //slope parameters
vector[NDim] subjAbil[Nsubj];      //person parameter matrix
real muDiff;
real <lower=0> sigmaDiff;
real muDisc;
real <lower=0> sigmaDisc;

vector [NDim] muAbil;
cov_matrix [NDim] SigmaSubjAbil;
}
```

```
transformed parameters {
matrix[Ntotal, NAnsK] pMat;
vector[3] Disc_vector;
matrix[Ntotal, NAnsK] pAns;
for ( rowIdx in 1:Ntotal ) {
  pMat[rowIdx,1]=1;
}
for ( rowIdx in 1:Ntotal ) {
  for (AnsKx in 2:NAnsK){
    matrix[1,NDim] v1;
    v1=block(itemDisc,itemID[rowIdx],1,1,NDim);
    Disc_vector=to_vector(v1);
    pMat[rowIdx,AnsKx]=pMat[rowIdx,AnsKx-
1]*exp(Disc_vector'*subjAbil[subjID[rowIdx]]+itemDiff[itemID[rowIdx],AnsKx-1]);
  }
}
for ( rowIdx in 1:Ntotal ) {
  for (AnsKx in 1:NAnsK){
    pAns[rowIdx,AnsKx]=pMat[rowIdx,AnsKx]/sum(block(pMat,rowIdx,1, 1, NAnsK));
  }
}
}
```

```
model {
//the hyperpriors
muDiff ~ normal (0,10^6);
sigmaDiff ~ gamma(1,1);
muDisc ~ normal (0,10^6);
sigmaDisc ~ gamma(1,1);
muAbil  ~ multi_normal (zero3,10^6*W);
SigmaSubjAbil ~ inv_wishart(4.0, W[,]);
// the priors
to_vector(itemDiff) ~ normal(muDiff, sigmaDiff);
to_vector(itemDisc) ~ lognormal(muDisc, sigmaDisc);
subjAbil ~ multi_normal(muAbil,SigmaSubjAbil);
// the likelihood
for ( rowIdx in 1:Ntotal ) {
  y[rowIdx] ~ categorical_logit(to_vector(block(pAns,rowIdx,1, 1, NAnsK)'));
}
}
"
writeLines( modelString , con="hirt.stan" )
library("rstan")
#rstan_options(auto_write = TRUE)
#options(mc.cores = parallel::detectCores())
```

```
library("parallel")
memory.limit(56000)
hirt.model<- stan(file  = "hirt.stan", data = dataList ,chains = 0)
Nchains <- 3
Niter <- 15000/3
initsList <-list()
set.seed(123456)
for (i in 1:3){
  initsAbil = matrix(rnorm(500*3),nrow=500)
  initsDisc = matrix(exp(rnorm(30*3)),nrow=30)
  initsDiff = matrix(rnorm(30*4),nrow=30)
  thisList=list(subjAbil=initsAbil,itemDiff=initsDiff,itemDgisc=initsDisc,.RNG.name=
"base::Super-Duper",
                .RNG.seed=123456+length(initsList))
  initsList[[length(initsList)+1]] <- thisList
}



t_start <- proc.time()[3]
fit<-stan(fit = hirt.model, data =dataList,pars=c("subjAbil", "itemDiff", "itemDisc"),
        chains = Nchains, iter=Niter,thin=1,init=initsList)
t_end <- proc.time()[3]
```

```
t_elapsed <- t_end - t_start

(time <- t_elapsed / Nchains / (Niter/2))

NUTSH <- As.mcmc.list(fit)

save(NUTSH , file="NUTSH.Rdata")

outNUTSH <-  as.data.frame(as.matrix(fit))

write.csv(outNUTSH, "outNUTSH.csv")

fit_summary <- summary(fit)

fitNUTSH <- fit_summary$summary

write.csv(fitNUTSH, "fitNUTSH)
```

Appendix C

Rationale for NOT Needing IRB Review

This study is to explore how two Bayesian estimation methods: Gibbs Sampler and Hamiltonian Monte Carlo-NO-U-TURN-SAMPLER perform in parameter estimation for a multidimensional partial credit item response model. A set of simulated datasets are created according to statistical simulation procedures, and the parameters are estimated by using the two Bayesian estimation methods. Finally, the Bias and Root Mean Square Errors are calculated to evaluate the performance of the two algorithms in different simulated conditions.

The study involves no living individuals, no interventions, no interactions (through surveys, interviews, tests, and observations), no identifiable private information, no existing data, no collaboration with other institutions, no engagement of University of Denver. Therefore, it does not need IRB review.