ABSTRACT

Title: FACTOR MIXTURE MODELS WITH ORDERED CATEGORICAL OUTCOMES: THE MATHEMATICAL RELATION TO MIXTURE ITEM RESPONSE THEORY MODELS AND A COMPARISON OF MAXIMUM LIKELIHOOD AND BAYESIAN MODEL PARAMETER ESTIMATION METHODS

Xiaodong Hou, Doctor of Philosophy, 2011

Directed By: Professor Gregory R. Hancock, Department of Measurement, Statistics and Evaluation

A factor mixture model (FMM) is a hybrid of latent class analysis and factor analysis modeling techniques. It can be used to investigate group differences in the absence of known class membership. The current study investigates the relation between FMMs and mixture item response theory (IRT) models. A formal proof of the mathematical equivalence between mixture graded-response models and FMMs with ordered categorical outcomes is presented and conversion formulas between the parameters of the two types of models are provided. More importantly, the current study conducts a Monte Carlo simulation study to compare Bayesian estimation with three different priors and maximum likelihood (ML) approach in fitting FMMs. Parameter recovery and classification accuracy are evaluated and compared. Besides the estimation method, the sample size, the number of outcome indicators, and the

magnitude of factor loadings are manipulated in the simulation. It is found that in general that ML and Bayesian estimation with weakly informative priors perform well with a small sample size, and that all estimation methods perform well with a large sample size. The results of this simulation also have implications for mixture IRT models based on its relation to FMMs.

FACTOR MIXTURE MODELS WITH ORDERED CATEGORICAL OUTCOMES: THE MATHEMATICAL RELATION TO MIXTURE ITEM RESPONSE THEORY MODELS AND A COMPARISON OF MAXIMUM LIKELIHOOD AND BAYESIAN MODEL PARAMETER ESTIMATION METHODS

By

Xiaodong Hou

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor Gregory R. Hancock, Chair
Professor Edward L. Fink
Professor Hong Jiao
Professor Robert W. Lissitz
Professor George Macready

Dedication

This dissertation is dedicated to my beloved family: my parents, my husband and my son. Your unconditional love and support are always my sources of progress, momentum, and happiness. Your beautiful smiles could turn a rainy day into sunshine.

Acknowledgements

I have been lucky and grateful during my graduate study. During this journey, I owe immense debt to so many people who have given me guidance, support, encouragement and opportunities. They made what I have achieved possible.

My special gratitude is to my advisor and committee chair Dr. Gregory R. Hancock who is always full of brilliant ideas and wisdom to share and who has been giving me tremendous and invaluable guidance and support in every single possible aspect. His mentorship was paramount throughout my graduate study and preparation for my future career. I hope I never stop learning from him. Having such a fabulous advisor is a honor and gift for me. My sincere and faithful gratitude for him is beyond any words.

My heart-felt thanks also go to my other committee members Dr. Robert W. Lissitz, Dr. Hong Jiao, Dr. George Macready and Dr. Edward L. Fink for their insightful suggestions and constructive comments. I would like to thank Dr. Lissitz for his important mentorship and support not only in my dissertation, but also in our work at the Maryland Assessment Research Center for Education Success. I am grateful for the opportunities he has given me. I would like to thank Dr. Jiao for her sharing her expertise and giving me invaluable inputs and help. I would like to thank Dr. Macready for his strong support and sharing his expertise with me. I would like to thank Dr. Fink for his important contribution to my dissertation and for giving me an enjoyable learning experience outside our department. In addition, I want thank

Mplus support group for their prompt support, especially suggestions from Drs. Muthén and Muthén.

My genuine gratitude also goes to Dr. William D. Schafer. I considerably benefited from his guidance, trust and the opportunities he has provided for me. I thank Dr. Robert Mislevy for his insights and the opportunity of working in his funded project. I thank Dr. David Paulson, Dr. Deborah Harris and Dr. JP Kim who awarded me summer internships. I thank Dr. Mitchell Dayton, Dr. Amy Hendrickson, Dr. Andre Rupp, and Dr. Jeffrey Harring for their tutoring during my study at University of Maryland. I thank Dr. David Miller, Dr. James Algina and Dr. Walter Leite for their mentorship during my study in Florida. In addition, I am grateful for the help from all my friends. I cherish our friendship and time we spent together during this journey.

Finally, my most important and deepest thanks go to my beloved family for all their unconditional love, support, sacrifice and faith in me. Thank you so much: my parents, my husband, and my child. I love you all with every beat of my heart.

TABLES OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

Chapter 1: Introduction 1
1.1 General Introduction to Latent Variable Factor Mixture Modeling 1
1.2 Contributions of this Dissertation Study 4
1.3 Brief Summary of the Following Chapters 7

Chapter 2: Theoretical Background 8
2.1 Theoretical Background for FMMs 8
2.1.1 Latent Class Analysis/Finite Mixture Modeling 8
2.1.2 Factor Analysis 9
2.1.3 FMMs with Continuous Outcomes 11
2.1.3 FMMs with Dichotomous Outcomes 12
2.2 Relationship between FMMs and Mixture IRT Models 12
2.3 Estimation Methods for FMMs/Mixture IRT Models 17
2.3.1 ML Estimation 17
2.3.2 Bayesian Inference 19
2.4 Issues in Fitting FMMs/Mixture IRT Models 24
2.4.1 Model Identification 24
2.4.3 Challenges 26
2.5 Software Packages 29

Chapter 3: Method 30
3.1 Data Generating Models 30
3.1.1 Manipulated Factors in the Simulation Study 33
3.2 Data Analysis Models 42
3.3 Evaluation of Estimation Outcomes 43
3.3.1 Parameter Recovery 43

Chapter 4: Results 45
4.1 Parameter Recovery 45
4.1.1 Recovery of Loading Parameters 51
4.1.2 Recovery of Class 1 Threshold Parameters 58
4.1.3 Recovery of Class 2 Threshold Parameters 63
4.2 Classification Accuracy 68
4.3 Convergence Rates 72

Chapter 5: Conclusion and Discussion 73
5.1 Summary of the Simulation Results 73
5.2 Future Study 77

Appendix A 80

Appendix B 83

References 119

LIST OF TABLES

Table 3-1 Priors for loadings and thresholds and corresponding 90% & 95% limit……..…34

Table 3-2 Simulation conditions……..36

Table 3-3 Summary of Monte Carlo population specifications for FMMs with Binary outcome……...……...……....38

Table 3-4 FMM data generating specification with 8 items and loading of 0.8……..39

Table 3-5 FMM data generating specification with 8 items and loading of 0.4……39

Table 3-6 FMM data generating specification with 30 items and loading of 0.8…..40

Table 3-7 FMM data generating specification with 30 items and loading of 0.4…..41

Table 4-1 Combination of manipulated factors……46

Table 4-2 One-way ANOVAs for impact of estimation method on four levels of sample size on relative bias of loadings…….……...57

Table 4-3 Average percentage of individuals assigned to the correct latent class….69

Table 4-4 Convergence rate……...72

Table A-1 Bias, relative bias and standard error of loading estimates……...80

Table A-2 Bias, relative bias and standard error of class 1 threshold estimates…….81

Table A-3 Bias, relative bias and standard error of class 2 threshold estimates…….82

LIST OF FIGURES

Figure 2-1 Relation of priors, likelihood and posterior distributions.……21
Figure 3-1 FMM in the simulation study……...31
Figure 4-1 Bias of loading ……..……..……..46
Figure 4-2 Bias of class 1 threshold ……..……..……...47
Figure 4-3 Bias of class 2 threshold ……..……..……...47
Figure 4-4 *SE* of loading ……...……..……49
Figure 4-5 *SE* of class 1 threshold …….....……...49
Figure 4-6 *SE* of class 2 threshold …….....50
Figure 4-7 Relative bias of loading for each item when N = 500, loading = 0.8 and number of items = 30……..……..…....53
Figure 4-8 Relative bias of loading for each item when N = 1000, loading = 0.8 and number of items = 30……..……..…....54
Figure 4-9 Relative bias of loading for each item when N = 2000, loading = 0.8 and number of items = 30……...……..……..…….....55
Figure 4-10 Relative bias of loading for each item when N = 5000, loading = 0.8 and number of items =30……..……..…....56
Figure 4-11 Relative bias of class 1 threshold for each item when N = 500, loading = 0.8 and number of items = 30……..……..…....59
Figure 4-12 Relative bias of class 1 threshold for each item when N = 1000, loading = 0.8 and number of items = 30……..……..…....60
Figure 4-13 Relative bias of class 1 threshold for each item when N = 2000, loading = 0.8 and number of items = 30……...……..…....61
Figure 4-14 Relative bias of class 1 threshold for each item when N = 5000, loading = 0.8 and number of items = 30..……..…....62
Figure 4-15 Relative bias of class 2 threshold for each item when N = 500, loading = 0.8 and number of items = 30……..……..…....64
Figure 4-16 Relative bias of class 2 threshold for each item when N = 1000, loading = 0.8 and number of items = 30…..……..…....65
Figure 4-17 Relative bias of class 2 threshold for each item when N = 2000, loading = 0.8 and number of items = 30……..…....66
Figure 4-18 Relative bias of class 2 threshold for each item when N = 5000, loading = 0.8 and number of items = 30……..…....67
Figure 4-19 Percentage of correct classification……..…....69
Figure 4-20 Scatter plots of posterior probabilities of belonging to the correct latent classes……..…....71
Figure B-1 Relative bias of loading for each item when N = 500, loading = 0.8 and number of items = 8…..……..…....83
Figure B-2 Relative bias of loading for each item when N = 1000, loading = 0.8 and number of items = 8……..……..…....84
Figure B-3 Relative bias of loading for each item when N = 2000, loading = 0.8 and number of items = 8……...……..…....85
Figure B-4 Relative bias of loading for each item when N = 5000, loading = 0.8 and number of items = 8……...……..…....86
Figure B-5 Relative bias of loading for each item when N = 500, loading = 0.4 and

number of items = 8……..…….....87
Figure B-6 Relative bias of loading for each item when N = 1000, loading = 0.4 and number of items = 8……...…….....88
Figure B-7 Relative bias of loading for each item when N = 2000, loading = 0.4 and number of items = 8….…….....89
Figure B-8 Relative bias of loading for each item when N = 5000, loading = 0.4 and number of items = 8….…….....90
Figure B-9 Relative bias of class 1 threshold for each item when N = 500, loading = 0.8 and number of items = 8……..……....91
Figure B-10 Relative bias of class 1 threshold for each item when N = 1000, loading = 0.8 and number of items = 8……..…….....92
Figure B-11 Relative bias of class 1 threshold for each item when N = 2000, loading = 0.8 and number of items = 8…….....93
Figure B-12 Relative bias of class 1 threshold for each item when N = 5000, loading = 0.8 and number of items = 8…….....94
Figure B-13 Relative bias of class 1 threshold for each item when N = 500, loading = 0.4 and number of items = 8……..…….....95
Figure B-14 Relative bias of class 1 threshold for each item when N = 1000, loading = 0.4 and number of items = 8…….....96
Figure B-15 Relative bias of class 1 threshold for each item when N = 2000, loading = 0.4 and number of items = 8……...97
Figure B-16 Relative bias of class 1 threshold for each item when N = 5000, loading = 0.4 and number of items = 8……..98
Figure B-17 Relative bias of class 2 threshold for each item when N = 500, loading = 0.8 and number of items = 8…….....99
Figure B-18 Relative bias of class 2 threshold for each item when N = 1000, loading = 0.8 and number of items = 8…….....100
Figure B-19 Relative bias of class 2 threshold for each item when N = 2000, loading = 0.8 and number of items = 8……...101
Figure B-20 Relative bias of class 2 threshold for each item when N = 5000, loading = 0.8 and number of items = 8…….....102
Figure B-21 Relative bias of class 2 threshold for each item when N = 500, loading = 0.4 and number of items = 8……..…….....103
Figure B-22 Relative bias of class 2 threshold for each item when N = 1000, loading = 0.4 and number of items = 8…….....104
Figure B-23 Relative bias of class 2 threshold for each item when N = 2000, loading = 0.4 and number of items = 8…….....105
Figure B-24 Relative bias of class 2 threshold for each item when N = 5000, loading = 0.4 and number of items = 8……106
Figure B-25 Relative bias of loading for each item when N = 500, loading = 0.4 and number of items = 30……....107
Figure B-26 Relative bias of loading for each item when N = 1000, loading = 0.4 and number of items = 30…….....108
Figure B-27 Relative bias of loading for each item when N = 2000, loading = 0.4 and number of items = 30...…….....109
Figure B-28 Relative bias of loading for each item when N = 5000, loading = 0.4 and

number of items = 30……..110
Figure B-29 Relative bias of class 1 threshold for each item when N = 500, loading = 0.4 and number of items = 30……..111
Figure B-30 Relative bias of class 1 threshold for each item when N = 1000, loading = 0.4 and number of items = 30……..112
Figure B-31 Relative bias of class 1 threshold for each item when N = 2000, loading = 0.4 and number of items = 30……..113
Figure B-32 Relative bias of class 1 threshold for each item when N = 5000, loading = 0.4 and number of items = 30……..114
Figure B-33 Relative bias of class 2 threshold for each item when N = 500, loading = 0.4 and number of items = 30……..115
Figure B-34 Relative bias of class 2 threshold for each item when N = 1000, loading = 0.4 and number of items = 30……..116
Figure B-35 Relative bias of class 2 threshold for each item when N = 2000, loading = 0.4 and number of items = 30……..117
Figure B-36 Relative bias of class 2 threshold for each item when N = 5000, loading = 0.4 and number of items = 30……..118

Chapter 1: Introduction

This chapter gives an introduction to latent variable factor mixture modeling, followed by contributions of this dissertation study and summary of the next chapters.

1.1 General Introduction to Latent Variable Factor Mixture Modeling

Latent variable (LV) modeling techniques have come to be widely used, especially in psychometrics and the behavioral sciences. To make inferences regarding a LV, observed outcome variables are usually used as imperfect and indirect observations of the LV. Both LVs and observed variables can be continuous or categorical.

The populations investigated in the social sciences are often heterogeneous, containing subgroups that may be known or unknown to the researchers. When group membership is known, such as those of gender, race, or a manipulated factor in an experimental design, multigroup latent variable modeling techniques may be used. For the examples of techniques that are able to address research questions on population differences in a latent variable system with known class membership such as multiple indicator multiple cause modeling and structured means modeling, see Hancock (2004) and Hancock, Lawrence, and Nevitt (2000). However, group membership is not always known beforehand and must be inferred from the data. In this situation, such groups are often referred to as *latent classes* because they cannot be directly observed, and special techniques for the analysis of unobserved heterogeneity in a population need to be used.

Latent class analysis (LCA) is popular among those techniques that can analyze unobserved population heterogeneity as reflected in distinct patterns of responses to measured items. However, traditional LCA can be problematic when subjects don't behave homogeneously within their estimated latent class or when subjects do not have the same conditional item probability within a class. On the other hand, factor analysis (FA) techniques can be used to investigate *m* unobserved continuous latent constructs that are believed to have causal influence on the *v* observed variables and account for the covariance among the *v* measured variables (where $m \leq v$) by seeing if the measured variables or items group together on continuous latent factors (Gorsuch, 1983). Both LCA and FA are approaches to explore or test hypotheses about the latent structures among the observed indictors (McCutcheon, 1987). However, LCA provides classification of subjects by identifying categorical latent variable(s) from a set of observed outcomes (Green, 1951, 1952), whereas FA characterizes one or more continuous or dimensional latent constructs from a group of observed indicators (Lazarsfeld & Henry, 1968).

By incorporating FA into LCA, the limitations of the two modeling techniques discussed above can be addressed. The combination of the two models is often called a *factor mixture model* (*FMM*; Muthén, 2006). An FMM has one or multiple categorical latent variables and one or more continuous latent variables. The single categorical latent variable is used to model latent class membership. Within each class, an FA model is specified by imposing a mean vector and covariance matrix of the observed variables (Lubke & Muthén, 2005).

In the behavioral sciences, observed response variables are often noncontinuous, possibly being dichotomous, ordinal, counts, or durations. The current study focuses on dichotomous and polytomous observed outcomes that are very common in the social sciences, particularly in testing and assessment.

When outcome variables are ordered and categorical, FA models are mathematically equivalent to grade response item response theory (IRT) models (see, e.g., Kamata & Bauer, 2008; Takane & Deleeuw, 1987); hence, one would assume FMMs to be mathematically equivalent to mixtures of IRT models (see, e.g., Clark et al., 2010; Masyn & Henderson, 2010; Muthén & Asparouhov, 2006). Mixture IRT models use a hybrid of the latent class model and IRT model and also have a single categorical latent variable and one or more continuous latent trait or factor variables (Dayton & Macready, 2007; Mislevy & Verhelst, 1990; Rost, 1990; von Davier & Rost, 2006). Mixture IRT provides an alternative approach for modeling categorical outcomes in the presence of unknown population heterogeneity, and within each latent class an IRT model is directly expressed in the form of a conditional probability of obtaining a score on the observed measure given the person's location usually on a unidimensional latent trait space (Millsap & Yun-Tein, 2004) and item parameters. The categorical latent variable plays the same role as the one in an FMM in modeling the unknown population heterogeneity, and within each subpopulation or class the same IRT model is assumed to hold (Hou & Hancock, 2010).

The popularity of FMMs is increasing mainly because they offer a way of investigating group differences in the absence of known class membership. FMMs have been suggested for detecting population heterogeneity in the behavioral sciences

and related areas, such as psychopathology, alcohol dependence, genetics, attention-deficit or hyperactivity disorder, twin heritability, aggressive behavior, social desirability, and bias of test items (Gagné, 2006; Kim & Muthén, 2009; Leite & Cooper, 2010; Lubke & Muthén, 2005; Lubke et al., 2007; Mann, 2009; McLachlan, Do, & Ambroise, 2004; Muthén, 2006; Muthén, Asparouhov, & Rebollo, 2006). Meanwhile, mixture IRT models have found their applications in educational measurement such as detecting latent groups that use different strategies to solve test items, modeling test speededness, helping maintain scale stability in the presence of test speededness, and detecting differential functioning in items and testlets (Bolt, Cohen, & Wollack, 2001, 2002; Cohen & Bolt, 2005; Cohen, Gregg, & Deng, 2005; Dai, 2009; Jiao et al., 2010; Mislevy & Verhelst, 1990; Samuelsen, 2005).

1.2 Contributions of this Dissertation

Based on the mathematical equivalence between FA with ordered categorical outcomes and IRT including 2PL and IRT Graded Response models (Kamata & Bauer, 2008; Muthén & Asparouhov, 2002; Takane & Deleeuw, 1987), both the FMMs and mixture IRT models, which cover the same types of categorical data, are assumed to be mathematically equivalent, as has been alluded to by several researchers (Clark et al., 2010; Masyn & Henderson, 2010; Muthén & Asparouhov, 2006). In the current study a formal proof of the mathematical equivalence of the FMMs and IRT Graded Response models is presented and conversion formulas between the parameters of the two types of models are provided for the conditions of multiple latent factors or traits and polytomous ordinal outcomes, of which the

unidimensional model with binary outcomes is a special case. It is hoped that this study provides a generic framework for factor mixture and mixture 2PL and mixture Graded Response modelers, and that this study promotes diffusion of the techniques in the social sciences.

More importantly, researchers who use FMMs or mixture IRT models have access to two estimation options: Bayesian and maximum likelihood (ML). Bayesian and ML analyses of the same data can yield different estimates (Browne & Draper, 2006), and the FMM or mixture IRT practitioners may wonder which estimates should be reported. In addition, Markov chain Monte Carlo (MCMC) is popular among mixture IRT modelers using WinBUGS (Spiegelhalter, Thomas, & Best, 2000), whereras marginal ML is predominantly used among FMM modelers using Mplus (Muthén & Muthén, 2010).

In the current study, we use a Monte Carlo simulation to compare Bayesian and ML approaches for fitting FMMs. The results of this simulation may also apply to and have implications for mixture IRT modeling based on the relation between FMMs and mixture IRT models. The likelihood approach examined in this study is the widely-used Marginal ML with the Expectation-Maximization (EM) algorithm, whereas the Bayesian method in this study uses the MCMC algorithm with Gibbs sampling and several sets of prior distributions.

Although there has been some research comparing ML and Bayesian approaches for the analysis of latent variable models (e.g., Browne & Draper, 2006), in the area of mixture modeling comparisons of the two estimation methods have just begun. Mixture IRT modeling is often done using WinBUGS (Spiegelhalter, Thomas,

& Best, 2000) with a Bayesian approach (e.g. Cohen & Bolt, 2005) and researchers usually choose diffuse priors, though for some models informative priors need to be used to achieve convergence. For factor mixture modeling, there was no software available for researchers and practitioners to fit models via Bayesian estimation until the recent release of Mplus 6 (Muthén & Muthén, 1998-2010). There has been a lack of research that compares the ML and Bayes estimation approaches in mixture modeling framework and explores the influence of priors on the estimates of FMMs. This dissertation will use simulation methods to examine the relative performance of Bayesian and ML estimation methods. Specifically, this simulation study is designed to answer the following questions:

(1) As sample size increases, which estimation method provides better parameter recovery and classification accuracy in fitting FMMs?

(2) As sample size increases, which prior performs better in fitting FMMs in Bayesian estimation with respect to parameter recovery and classification? As sample size increases, is the effect of the priors on Bayesian estimation negligible, rendering it comparable to ML in terms of parameter recovery and classification?

(3) As the magnitude of loadings increases, do the parameter recovery and classification improve in either or both estimation methods?

(4) As the magnitude of loadings increases, which prior performs best in fitting FMMs in Bayesian estimation in terms of the parameter recovery and classification?

(5) As the number of binary indicators increases, do the parameter recovery and classification improve in either or both estimation methods?

(6) As the number of binary indicators increases, which prior performs best in fitting FMMs in Bayesian estimation in terms of the parameter recovery and classification?

1.3 Brief Summary of the Following Chapters

The second chapter presents a review of literature for the FMMs with two estimation methods, including choice of priors in Bayesian estimation. This chapter also provides a formal proof of the mathematical equivalence between the FMMs and mixture IRT Graded Response models. Chapter 3 proposes a systematic simulation design to compare the performance of the two estimation methods in recovering the underlying latent structure in the factor mixture models. Chapter 4 provides the results of the simulation study. Finally, Chapter 5 is a discussion of the findings and limitations of the current study followed by the description of potential future research extensions and recommendations for practitioners.

Chapter 2: Theoretical Background

This chapter details the theoretical background of the current study, which includes the theoretical background of FMMs, the mathematical relation between FMMs and mixture IRT models, estimation methods, issues in fitting FMMs and mixture IRT models and available software packages.

2.1 Theoretical Background for FMMs

2.1.1 Latent Class Analysis/Finite Mixture Modeling

Latent class analysis (LCA), also known as finite mixture modeling, was first developed by Lazarsfeld and Henry (1968), and it is aimed at identifying unobserved subgroups of a heterogeneous population. The observed response in LCA can be categorical, continuous, a count, or censored data. The latent variable is often categorical, representing membership in the subgroups or latent classes.

Conditional item probabilities and class probabilities are the two important model parameters in an LCA with binary outcomes. Consider such a LCA model with latent class variable c with K classes ($c = k$; $k = 1, \ldots, K$) and the vector-valued response $\mathbf{X} = \{x_i\}$ for item $i = 1,\ldots, v$; the marginal item probability for $x_i = 1$ is

$$P(x_i = 1) = \sum_{k=1}^{K} P(c = k)P(x_i = 1 \mid c = k). \qquad (2\text{-}1)$$

The joint probability of any specified response vector $\mathbf{X} = \{x_i\}$ across all of the items may be expressed as

$$P(x_1,...,x_v) = \sum_{k=1}^{K} P(c = k)P(x_1 \mid c = k)\cdots P(x_v \mid c = k). \qquad (2\text{-}2)$$

The posterior probabilities, the estimated class probabilities for each subject are

$$P(c = k \mid x_1,...,x_v) = \frac{P(c = k)P(x_1 \mid c = k)\cdots P(x_v \mid c = k)}{P(x_1,...,x_v)}. \qquad (2\text{-}3)$$

Each subject may have nonzero probability values for many classes. There are different ways of classifying respondents. Taking class probabilities as an example, if the probability of being in a class is the highest among the nonzero probability values, a subject is assigned to that class and then assumed to be in only that class.

Let π_k indicate the proportion of a mixture component *k*. It is assumed that $\sum_{k=1}^{K} \pi_k = 1$ for $k = 1,\ldots, K$ classes. In addition, the correlation among the responses is assumed to be accounted only by the latent class variable and that subjects are assumed to behave homogenously within their estimated latent class. The first assumption often results in adding more unnecessary latent classes that only account for residual correlations between a small number of items (Clark, 2010). This second assumption is often problematic when subjects have continuous or dimensional latent constructs within a class (Muthén & Asparouhov, 2006).

2.1.2 Factor Analysis

Factor analysis (FA) is a widely used modeling technique, which can be used to explore the relationship between latent and observed variables in exploratory FA or test hypotheses about the particular latent structures among the observed indictors in confirmatory FA. Specifically, it investigates *m* unobserved latent constructs that are

believed to have causal influence on *v* observed variables and account for the covariance among *v* measured items responses (where $m \leq v$) by seeing if items group together on the continuous latent constructs called factors (Gorsuch, 1983). The general forms of FA with continuous and categorical outcome variables are given below (for more detail see Brown, 2006; Jöreskog, 1969). A general factor analysis model with *v* continuous outcomes can be expressed as

$$\mathbf{Y} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon}, \tag{2-4}$$

where $\mathbf{Y}$ is the $v \times 1$ vector of observed outcomes only if the observed variables are continuous, otherwise $\mathbf{Y}$ is assumed to be a continuous underlying latent response (see details in the next paragraph); $\mathbf{\Lambda}$ is a $v \times m$ matrix of loadings λ, m is the number of factors; $\boldsymbol{\xi}$ is a $m \times 1$ factor score vector; and $\boldsymbol{\varepsilon}$ is a $v \times 1$ residual vector. The residuals are typically assumed to be independently normally distributed.

In an FA with ordered categorical outcomes, it is assumed that there is a continuous underlying latent response $\mathbf{Y}$, which is a combination of the common factor and item-specific residual. Taking dichotomous outcomes as an example, the categorization of $\mathbf{Y}$ into observed dichotomous variable $\mathbf{X}$ can be expressed as the following threshold model for item *i*:

$$X = \begin{cases} 1 & if \quad y_i \geq \tau \\ 0 & if \quad y_i < \tau \end{cases} \tag{2-5}$$

where τ is the threshold for item *i*. This threshold model describes the nonlinear relationship between observed $\mathbf{X}$ and underlying latent response $\mathbf{Y}$.

2.1.3 FMMs with Continuous Outcomes

FMM is a hybrid of LCA and FA models (Muthén, 2006) and has one or more categorical latent variables and multiple continuous latent variables. The categorical latent variables are used to model latent class membership. Within each class, an FA model is specified by imposing a mean vector and covariance matrix of the observed variables (Lubke & Muthén, 2005).

A general FMM with continuous outcomes $\mathbf{Y} = \{y_{iks}\}$ for person s ($s = 1,\ldots, S$) can be formulated as:

$$\mathbf{Y} = \boldsymbol{\Lambda}_{ik}\boldsymbol{\xi}_{ks} + \boldsymbol{\varepsilon}_{ks} \quad , \tag{2-6}$$

where for each item within each class, $\mathbf{Y}$ is the $v \times 1$ vector of observed outcomes only if the observed variables are continuous, otherwise $\mathbf{Y}$ is assumed to be a $v \times 1$ vector of continuous underlying latent response (see the discussion in factor mixture models with categorical outcomes); $\boldsymbol{\Lambda}_{ik}$ is a $v \times m$ matrix of loadings, m is the number of factors; $\boldsymbol{\xi}_{ks}$ is a $m \times 1$ factor score vector; and $\boldsymbol{\varepsilon}_{ks}$ is a $v \times 1$ residual vector. The residuals are typically assumed to be independently normally distributed, although a logistic distribution can also be considered. In addition, it is often assumed according to the factor model that observed outcome variables $\mathbf{Y}$ within each mixture component follow a class-specific multivariate normal distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, with $\boldsymbol{\mu}_k = \boldsymbol{\Lambda}_{ik}\boldsymbol{\alpha}_k$ and $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_{ik}\boldsymbol{\Theta}_k\boldsymbol{\Lambda}'_{ik} + \boldsymbol{\Psi}_k$, where $\boldsymbol{\alpha}_k$ is the vector of factor means, $\boldsymbol{\Theta}_k$ is the factor covariance matrix, and $\boldsymbol{\Psi}_k$ is residual variance matrix. $\boldsymbol{\Psi}_k$ can be assumed to be diagonal (i.e., reflecting local independence), although this restriction can be relaxed.

2.1.4 FMMs with Dichotomous Outcomes

In the FMM with an ordered categorical outcome variable, it is assumed that there is a continuous underlying latent response $\mathbf{Y} = \{y_{iks}\}$ for person *s*. The categorization of **Y** into observed dichotomous variable **X** can be expressed as the following threshold model for item *i*:

$$X = \begin{cases} 1 & \text{if} \quad y_{iks} \geq \tau_{ik} \\ 0 & \text{if} \quad y_{iks} < \tau_{ik} \end{cases} \tag{2-7}$$

where τ_{ik} is the threshold for item *i* in class *k*. In the (2-7), the observed dichotomous outcomes **X** are related to the unobserved **Y** through threshold τ_{ik}. For the description of the FMMs with polytomous outcome variables, see the next section on the relation between FMMs and mixture IRT models.

2.2 Relationship between FMMs and Mixture IRT Models

Based on the mathematical equivalence between FA with ordered categorical outcomes and IRT including 2PL and Graded Response models (Kamata & Bauer, 2008; Muthén & Asparouhov, 2002; Takane & Deleeuw, 1987), both these FMMs and mixture IRT models, which cover the same types of categorical data, are often assumed to be mathematically equivalent, as has been alluded to by several researchers (Clark, 2010; Masyn & Henderson, 2010; Muthén & Asparouhov, 2006). In this section, a formal proof of mathematical equivalence of multidimensional FMMs with polytomous outcome variables and multidimensional mixture grade-

response IRT models with polytomous outcome variables is provided, followed by the transformation formulas between the model parameters. The model with binary outcomes is a special case of the one with polytomous outcome. It is hoped that this study provides a generic framework for both types of models.

Suppose a random vector of response pattern to v ordered categorical items is $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_v)$ and each $\mathbf{X}_i$ has j categories, where $j = 1, \ldots, v_i$ and $i = 1, \ldots, v$, then the j-th element of $\mathbf{X}_i' = (x_{i(1)}, \ldots, x_{i(v_i)})$ is defined as

$$x_{i(j)} = \begin{cases} 1, \text{if response to item } i \text{ is in category } j, \\ 0, \text{otherwise,} \end{cases} \tag{2-8}$$

where it is assumed that $x_{i(j)}x_{i(t)} = 0$, for $j \neq t$ (i.e. j and t represent different categories) and $\sum_{j=1}^{v_i} x_{i(j)} = 1$.

In an FMM with polytomous outcomes, the marginal probability of $\mathbf{X}$ is expressed as

$$P(\mathbf{X} = X) = \int_R h(y_s)dy_s, \tag{2-9}$$

where $\mathbf{Y} = \boldsymbol{\Lambda}_{ik}\boldsymbol{\xi}_{ks} + \boldsymbol{\varepsilon}_{ks}$ (2-6) as shown before. It is assumed that $\boldsymbol{\xi}_{ks} \sim$ iid $MVN(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\varepsilon}_{ks} \sim$ iid $MVN(\mathbf{0}, \boldsymbol{\psi}_k)$, where $\boldsymbol{\psi}_k$ is assumed to be diagonal, and $\boldsymbol{\xi}_{ks}$ and $\boldsymbol{\varepsilon}_{ks}$ are independent of each other. Therefore, the marginal distribution of $\mathbf{Y}$ is

$$\mathbf{Y} \sim \text{iid } \boldsymbol{MVN}(\mathbf{0}, \boldsymbol{\Lambda}_{ik}\boldsymbol{\Lambda}_{ik}' + \boldsymbol{\psi}_k) \tag{2-10}$$

And the conditional distribution of Y given $\boldsymbol{\xi}$ is

$$\mathbf{Y} \mid \boldsymbol{\xi}_{ks} \sim \text{iid } MVN(\boldsymbol{\Lambda}_{ik}\boldsymbol{\xi}_{ks}, \boldsymbol{\psi}_k). \tag{2-11}$$

In addition, in equation (2-9), *R* is the multidimensional region of integration. *R* is defined as the direct product of intervals R_i (i = 1, ...,v), where $R_i = (\tau_{i(j-1)}, \tau_{i(j)})$ if $x_{i(j)} = 1$. Note that $\tau_{i(j)}$ is the category boundary between the (j-1)th and jth categories; $\tau_{i(0)} = -\infty$ and $\tau_{i(v_i)} = \infty$.

In FMMs, the joint distribution of latent class variable *c* and the vector-valued response **Y** can be built as the product of the marginal distribution of *c* and **Y** given class *K*:

$$f(c, \mathbf{Y}) = f(c)\, f(\mathbf{Y} \mid c = K), \tag{2-12}$$

where $f(\cdot)$ is a probability distribution and *K* is the number of mixture components. Because π_k indicates the proportion of the mixture component / class *k*, the model can be further formalized as

$$f(\mathbf{Y}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{Y}) \tag{2-13}$$

(Lubke & Neale, 2008; Lubke & Spies, 2008). Therefore, Equation (2-9) can be rewritten as

$$\begin{aligned} P(\mathbf{X} = X) &= \int_R \sum_{k=1}^{K} \pi_k h_k(y_{ks})\, dy_{ks} \\ &= \sum_{k=1}^{K} \pi_k \int_R h_k(y_{ks})\, dy_{ks} \\ &= \sum_{k=1}^{K} \pi_k \int_R \left(\int_\Xi f(y_{ks} \mid \xi_{ks}) g(\xi_{ks})\, d\xi_{ks} \right) dy_{ks} \\ &= \sum_{k=1}^{K} \pi_k \int_\Xi g(\xi_{ks}) \left(\int_R f(y_{ks} \mid \xi_{ks})\, dy_{ks} \right) d\xi_{ks}. \end{aligned} \tag{2-14}$$

Because of (2-11) and the definition of *R*, Equation (2-14) can be further written as

$$\mathrm{P}(\mathbf{X}=\mathrm{X})=\sum_{\mathrm{k}=1}^{\mathrm{K}}\pi_{\mathrm{k}}\int_{\Xi}\mathrm{g}(\xi_{\mathrm{ks}})\left(\prod_{\mathrm{i}}\int_{\mathrm{R_i}}\mathrm{f}\left(\mathrm{y_{iks}}\mid\xi_{\mathrm{ks}}\right)\mathrm{dy_{iks}}\right)\mathrm{d}\xi_{\mathrm{ks}}$$

$$=\sum_{\mathrm{k}=1}^{\mathrm{K}}\pi_{\mathrm{k}}\int_{\Xi}\mathrm{g}(\xi_{\mathrm{ks}})\prod_{\mathrm{i}}\prod_{\mathrm{j}}\left\{\Phi\left(\frac{\Lambda'_{\mathrm{ik}}\xi_{\mathrm{ks}}-\tau_{\mathrm{i(j-1)k}}}{\sqrt{\psi_{\mathrm{iks}}}}\right)-\Phi\left(\frac{\Lambda'_{\mathrm{ik}}\xi_{\mathrm{ks}}-\tau_{\mathrm{ijk}}}{\sqrt{\psi_{\mathrm{iks}}}}\right)\right\}^{\mathrm{x_{ij}}}\mathrm{d}\xi_{\mathrm{ks}}. \quad (2\text{-}15)$$

The graded-response model was proposed by Samejima (1969): the normal ogive model and the logistic model for graded response data (i.e. ordered categorical outcomes), which specifies the probability of a respondent of a given ability receiving a rating score *0,…,M*. The mixture graded-response model with ordered categorical outcomes can be written as

$$P(\mathbf{X}=X\mid\xi_{\mathrm{ks}})=P_{\mathrm{i(j-1)}}(\xi_{\mathrm{ks}})-P_{\mathrm{ij}}(\xi_{\mathrm{ks}}). \quad (2\text{-}16)$$

The marginal probability of $\mathbf{X} = X$ in the normal-ogive mixture IRT model for graded response can be specified as

$$P(\mathbf{X}=X)=\sum_{k=1}^{K}\pi_k P_k(\boldsymbol{X}_{ks}=X)$$

$$=\sum_{\mathrm{k}=1}^{\mathrm{K}}\pi_{\mathrm{k}}\int_{\Xi}\mathrm{P_k}(\mathbf{X}_{\mathrm{ks}}=\mathrm{X}\mid\xi_{\mathrm{ks}})\mathrm{g}(\xi_{\mathrm{ks}})\mathrm{d}\xi_{\mathrm{ks}}. \quad (2\text{-}17)$$

Based on local independence, (2-18) is further derived from Equation (2-16) and (2-17),

$$P(\mathbf{X}=X)=\sum_{\mathrm{k}=1}^{\mathrm{K}}\pi_{\mathrm{k}}\int_{\Xi}\left(\prod_{\mathrm{i}}\prod_{\mathrm{j}}(P_{\mathrm{i(j-1)}}(\xi_{\mathrm{ks}})-P_{\mathrm{ij}}(\xi_{\mathrm{ks}}))\right)^{\mathrm{x_{ij}}}\mathrm{g}(\xi_{\mathrm{ks}})\mathrm{d}\xi_{\mathrm{ks}}. \quad (2\text{-}18)$$

Let Φ be the standard normal cumulative distribution function, $\xi_{ks} \sim N(0, I)$, and *a* is the discrimination parameter and *b* the difficulty parameter, we get

$$\begin{aligned} P_{ij}(\xi_{ks}) &= \int_{-\infty}^{a'_{ik}(\xi_{ks}-b_{ijk})} \Phi(z)dz \\ &= \Phi(a'_{ik}(\xi_{ks} - b_{ijk})). \end{aligned} \tag{2-19}$$

Therefore,

$$\begin{aligned} &P(\mathbf{X} = X) \\ &= \sum_{k=1}^{K} \pi_k \int_{\Xi} g(\xi_{ks}) \prod_i \prod_j \left\{ \Phi\left[a'_{ik}(\xi_{ks} - b_{i(j-1)k})\right] - \Phi\left[a'_{ik}(\xi_{ks} - b_{ijk})\right] \right\}^{x_{ij}} d\xi_{ks}. \end{aligned} \tag{2-20}$$

Recall (2-15) is

$$\sum_{k=1}^{K} \pi_k \int_{\Xi} g(\xi_{ks}) \prod_i \prod_j \left\{ \Phi\left(\frac{\Lambda'_{ik}\xi_{ks} - \tau_{i(j-1)k}}{\sqrt{\psi_{iks}}} \right) - \Phi\left(\frac{\Lambda'_{ik}\xi_{ks} - \tau_{ijk}}{\sqrt{\psi_{iks}}} \right) \right\}^{x_{ij}} d\xi_{ks}$$

where

$$\boldsymbol{a_{ik}} = \frac{\boldsymbol{\lambda_{ik}}}{\sqrt{\boldsymbol{\psi_{iks}}}} \tag{2-21}$$

$$\boldsymbol{b_{ijk}} = \frac{\boldsymbol{\tau_{ijk}}}{\boldsymbol{\lambda_{ik}}}. \tag{2-22}$$

In summary, the item discrimination parameter is equal to the value of the factor loading parameters divided by the residual variance, and item difficulty parameters are associated with both factor loading and threshold parameters. Please note that discrimination parameters are always constrained positive in IRT, but corresponding loading parameters can be negative in factor analysis. The current proof assumes an ability testing context where all loadings are in one direction so that all loadings can be considered and constrained to be positive.

2.3 Estimation Methods for FMMs/mixture IRT Models

There are currently two estimation methods for fitting FMMs and three for fitting mixture IRT models. For FMMs, one estimation method is a classic frequentist approach, ML (marginal ML to be exact), and the other is Bayesian inference. ML is widely used in the social sciences, whereas Bayesian methods' popularity is growing with the use of the computational algorithm MCMC and the development of various sampling methods such as the Gibbs sampler (Geman & Geman, 1984) and the Metropolis-Hastings algorithm (Hastings, 1970). For mixture IRT models, besides marginal ML and Bayesian estimation discussed above, conditional ML is an alternative method for mixed Rasch models.

2.3.1 ML Estimation

ML estimation is the most popular method for estimating model parameters in the frequentist approach. In a frequentist approach, the first step is to hypothesize a statistical model to describe the data and to determine a probability distribution to model the response distribution. Suppose $y_1, y_2, \ldots, y_s$ are independent observations of a random variable **Y**; the joint probability of the vector **Y** given a set of model parameter **θ** is expressed as

$$\Pr(\mathbf{Y} \mid \theta) = \Pr(y_1 \mid \theta) \cdot \Pr(y_2 \mid \theta) \cdots \Pr(y_s \mid \theta). \tag{2-22}$$

Because we are interested in estimating the model parameters given the observed variable, we rewrite (2-23) and get a likelihood function as follows:

$$\Pr(\mathbf{Y}|\theta) \equiv L(\theta|y) = \prod_{s=1}^{S} \Pr(y_s \,|\theta), \tag{2-23}$$

where $\Pr(.)$ on the right hand side represents the probability distribution that is believed to generate the observed data (Lynch, 2007). To construct a likelihood function, an appropriate probability distribution $\Pr(.)$ needs to be determined.

In factor mixture modeling, if we assume that π_k indicates the proportion of a mixture component, and that $\mathbf{Y}$ within each mixture component follows a class-specific multivariate normal distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$; the likelihood function for a FMM can be expressed as

$$L(\boldsymbol{\theta} \mid y) = \prod_{s=1}^{S} \Pr(y_s \mid \pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{s=1}^{S} \sum_{k=1}^{K} \pi_k \frac{1}{\left(\sqrt{2\pi}\right)^v \sqrt{|\boldsymbol{\Sigma}_k|}} e^{-\frac{1}{2}(y_s-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(y_s-\boldsymbol{\mu}_k)} . \quad (2\text{-}24)$$

After establishing the likelihood function, the second step in ML estimation is to estimate model parameters given the observed data in the above likelihood function. By finding the parameter values to maximize the probability function, ML estimation obtains the values of model parameters that produce the distribution most likely to have resulted in the observed variables (i.e., that make the observed data most likely to have occurred). In ML estimation, instead of directly using the likelihood function, the logarithm of the likelihood function is usually applied because first, the log-likelihood is more convenient to work with by converting the repeated multiplication to repeated addition, and second, both functions reach maxima at the same point (Lynch, 2007). The log-likelihood function for a FMM can be expressed as

$$\boldsymbol{ln}\, L(\boldsymbol{\theta} \mid y) = \sum_{s=1}^{S} \ln \Pr(y_s \mid \pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{s=1}^{S} \ln \left(\sum_{k=1}^{K} \pi_k \frac{1}{(\sqrt{2\pi})^v \sqrt{|\boldsymbol{\Sigma}_k|}} e^{-\frac{1}{2}(y_s-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(y_s-\boldsymbol{\mu}_k)} \right). \quad (2\text{-}25)$$

Because both latent factor and latent class memberships are unobserved, there is no closed form solution for the parameter estimates. The expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) or some modification of EM is often used to obtain ML estimates (McLachlan & Krishnan, 2008). In this study, the statistical analysis via ML estimation with the EM algorithm with Quasi-Newton and Fisher Scoring acceleration is carried out in Mplus 6.1. EM algorithm can be used for finding ML parameter estimates in the presence of missing data. Specifically, the EM algorithm is used for finding estimates for unobserved latent variables and class membership in FMM estimation (Mann, 2009; Yung, 1997). The EM algorithm is an iterative procedure which alternates between an expectation (E) step and a maximization (M) step. In the E step, the expectation of the log-likelihood is computed using the initial or current estimates for the latent variables and latent class membership. In the M step, parameter estimates are calculated by maximizing the expected log-likelihood obtained in the E step. These parameter estimates are then used to determine the estimates for the latent variables and class membership in the next E step (Yung, 1997). Because the EM algorithm with a large amount of the missing data converges at an extremely slow rate, quasi-Newton and Fisher scoring algorithm are recommended and used (Lange, 1995; McLachlan & Krishnan, 2008; Muthén & Muthén, 2010; 1998-2010)) to accelerate convergence.

2.3.2 Bayesian Inference

Bayesian analysis has gained popularity with the use of the MCMC algorithm. In the Bayesian approach, a parameter is considered as a random variable instead of a

constant. The process of Bayesian inference is to first represent prior uncertainty about a parameter with a probability distribution and then to produce a posterior probability distribution with the current data in order to update and lessen the uncertainty about the parameter (Lee, 2007; Lynch, 2007).

In Bayesian inference, all knowledge and uncertainty about the unknown estimates are measured by probabilities. Consider the probabilities of events A and B, P(A) and P(B). The joint event A and B can be expressed in terms of conditional and marginal probabilities:

$$P(A, B) = P(A|B)\,P(B) = P(B|A)\,P(A). \tag{2-26}$$

Dividing by P (A), we get Bayes' Theorem:

$$\mathrm{P(B \mid A)} = \frac{\mathrm{P(A \mid B)P(B)}}{\mathrm{P(A)}}. \tag{2-27}$$

Let the data be A and the prior information be B, so thus we obtain

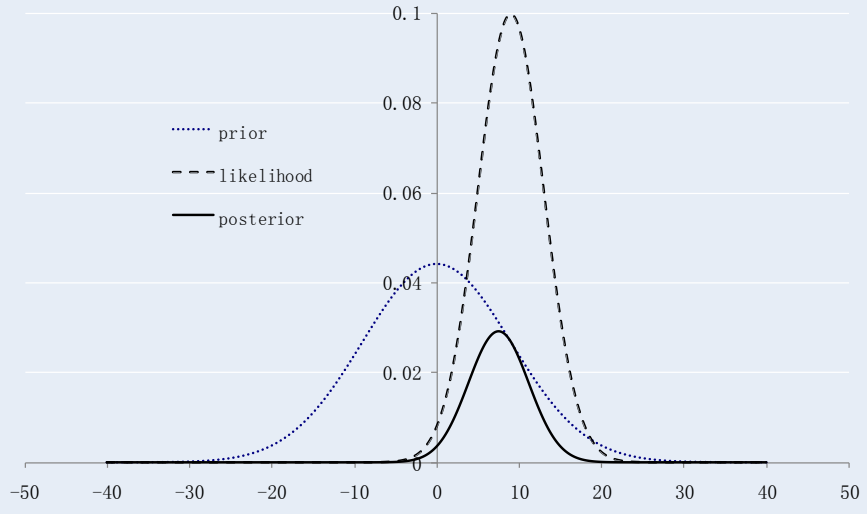
$$\mathrm{P(prior \mid data)} = \frac{\mathrm{P(data \mid prior)P(prior)}}{\mathrm{P(data)}}, \tag{2-28}$$

where *P(prior|data)* is referred to as a *posterior distribution*. *P(data|prior)* is the sampling distribution or likelihood of the data given the prior. The density *P(prior|data)* is the probability belief on prior after seeing the data. The factor *P(data)* does not depend on the prior so it does not need updating when iteratively finding the posterior and it can be considered to be a constant. Therefore, the posterior distribution can be rewritten as

$$Posterior \propto Likelihood \times Prior. \tag{2-29}$$

After observing new data from a study, knowledge of a prior is updated through Bayes' theorem. Figure 2-1 is an example that shows the relation among prior, likelihood and posterior distributions in Bayes' theorem.

Figure 2-1. Relation among prior, likelihood, and posterior distributions



The posterior distribution is possible to be analytically derived for models with closed forms. For complex statistical models without closed forms like FMMs, MCMC algorithms (Gamerman, 1997) have been widely used in which *Monte Carlo* means the random simulation process and *Markov Chain* refers to making random draws of parameter values from the posterior distribution given the previous set (Lynch, 2007). That is, the MCMC simulates samples from the posterior distribution and then makes inferences based on these posterior samples (Yuan & MacKinnon, 2009). Gibbs sampling is the most basic and widely used MCMC algorithm. It updates parameter values based on their conditional distributions. In other words, it samples a set of new parameter values at iteration *q* from the conditional distribution given the other parameter values at iteration *q*-1. Suppose $\boldsymbol{\theta}^q$ is a vector of model

parameters at iteration q ($q = 1,\ldots,Q$), the updated value of the parameter vector is obtained from the following iterative steps of Gibbs sampling:

1. Set starting values to the parameter vector when $q = 0$.
2. Let $q = q+1$.
3. Sample $\boldsymbol{\theta}_1^q \mid \boldsymbol{\theta}_2^{q-1},...,\boldsymbol{\theta}_{w-1}^{q-1}, \text{data}, \text{priors}.$
4. Sample $\boldsymbol{\theta}_2^q \mid \boldsymbol{\theta}_1^q,\boldsymbol{\theta}_3^{q-1}...,\boldsymbol{\theta}_{w-1}^{q-1}, \text{data}, \text{priors}.$

   …

w. Sample $\boldsymbol{\theta}_w^q \mid \boldsymbol{\theta}_1^q,\boldsymbol{\theta}_2^q,...,\boldsymbol{\theta}_{w-1}^q, \text{data}, \text{priors}.$

where w is the number of sets of parameters in the vector $\boldsymbol{\theta}$ (i.e. $\boldsymbol{\theta}^q = (\theta_1^q,...,\theta_w^q)'$). After step w in iteration q, the loop from step 2 to step w is repeated for iteration q+1. This iterative process should not stop until convergence is achieved, which produces a Markov chain of parameter values drawn from the posteriors. Such iterations are referred to as a *chain*. Different starting values and random seeds when making random draws should be used in different chains. Usually two or more chains are used in Bayesian estimation (Lynch, 2007; Muthén & Asparouhov, 2010).

The estimates of model parameters in an FMM can be obtained through the Q iterations. For example, the point estimate of the threshold τ_{ik} is given by

$$\hat{\tau}_{ik} = \frac{1}{Q}\sum_{q=1}^{Q}\tau_{ik}{}^{q}. \tag{2-30}$$

And the point estimate of loading λ is

$$\hat{\lambda}_{ik} = \frac{1}{Q}\sum_{q=1}^{Q}\lambda_{ik}{}^{q}. \tag{2-31}$$

2.3.2.1 Priors for Factor Mixture Analysis with Ordered Categorical Indicators

Categorical outcome variables provide relatively less information than continuous outcomes. When both the sample size and the number of indicators are small, the data will contribute limited information and the prior will play an important role in Bayesian estimation.

A prior can be noninformative, weakly informative, or informative. A *noninformative prior*, also called a diffuse or flat prior, has a large variance, which shows large uncertainty of the parameter value and is used when researchers have little prior information on the parameters of interest, or when they prefer that the inference is not affected by information outside of the study (Gelman et al., 2008; Muthén & Asparouhov, 2010). There is a great deal of literature on noninformative and default prior distributions (e.g., Kass & Wasserman, 1996; Yang & Berger, 1994). The noninformative normally distributed priors for loadings, or a log-normal prior, or more complicated priors are found in many studies (Asparouhov & Muthén, 2010b; Fox & Glas, 2001; Lee et al., 2010; Patz & Junker, 1999; Segawa et al., 2008; Song et al., 2009).

In contrast, an *informative prior* has a small variance, which reflects confident prior beliefs of a parameter and can be used when prior knowledge is available (Brown & Draper, 2006). Relatively speaking, the more informative the prior is, the less influence the likelihood has on the posterior distribution. The current study focuses on the priors between the noninformative and informative priors, called *weakly informative priors*, based on the characteristics of the probit regression used in fitting the FMMs with categorical outcomes.

Weakly informative priors were suggested by Gelman and his colleagues (2008) for logistic and probit regressions. As they pointed out, a weakly informative prior would be more appropriate for default use in a wide range of applications in logistic and probit regression because a relatively small change in the logit or probit scale corresponds to a relatively large change in probability. For example, a change of 5 units on the logistic scale means the change of 50% on the probability scale, and a change of 5 units on the probit scale indicates a change of 99% on the probability scale. Weakly informative normally-distributed priors were found to work well in Bayesian estimation in their study, though the preference is the *t* distribution and the Cauchy priors. Such priors as N (0, 1), N(0, 5), and N(0, 20) are also recommended and used for threshold and loading parameters on the logit scale by Asparouhov and Muthén (2010b) in their simulation study on factor analysis with binary indicators. It is also suggested that using completely noninformative priors for parameters on the probit scale may induce skewed priors on probability scale.

2.4 Issues in Fitting FMMs/Mixture IRT Models

2.4.1 Model Identification

Like FA models, FMMs also have the problem of indeterminacy. In order to identify the model, one needs to assign a metric and an origin for the factors (Mann, 2009). The metric can be given by fixing a factor loading to unity or factor variances to unity in each class. There are two ways to assign an origin for the factors. One is to set the factor means in each class to zero, and the other is to set the factor mean of one class to zero and constrain the thresholds associated with the referent indicators

to be equal across classes (Mann, 2009; Muthén, 2008; Yung, 1997). In the current study, the factor means and factor variances are set to zero and unity, respectively, for giving the origin and metric of the factor. All thresholds and loadings across classes are estimated. From an IRT perspective, a similar parameterization is often applied for identification purposes. For example, in the 2PL-IRT model, the model can be identified by adding two types of constraints. One is to fix the origin and metric of the latent trait, which is usually done by setting the mean and variance of the latent trait to be zero and one, respectively. The other is to impose constraints on the item parameters such as assuming $\prod_i a_i = 1$ and $\sum_i b_i = 0$ (where a_i and b_i are item discrimination and difficulty parameters) or setting one discrimination parameter to one and one difficulty parameter to zero (Fox & Glas, 2001; Skrondal & Rabe-Hesketh, 2005).

2.4.2 Confirmatory or Exploratory

FMM can take two possible factor measurement structures: confirmatory or exploratory. Because the current study is not aimed at exploring the structure of the data, a confirmatory structure is assumed, which means that information such as the number of factors and the relation between the factors and items are the prior knowledge before the analysis.

2.4.3 Challenges

2.4.3.1 Local Maximum Solutions

The issue of local maxima has been long known in the method of ML (Goodman, 1974), which finds the solutions by maximizing the log-likelihood function discussed above. The ML algorithm terminates whenever a small change of model parameter estimates decreases the log-likelihood. This aspect of ML makes terminations possible at the global maximum or at a local optimal solution. To avoid local maxima, multiple sets of starting values are important to use in fitting FMMs (Vermunt & Magidson, 2005). It is recommended that the highest log-likelihood value should be replicated in at least two final stage solutions, indicating that a good solution is obtained (Muthén & Muthén, 1998-2010), though we can never know for sure if we get global maximum solutions.

2.4.3.2 Convergence

Failure to converge to a stable solution is often an issue in mixture modeling in both ML and Bayesian estimation. A binary outcome makes convergence difficult, perhaps due to too little information (Muthén, 2010). In mixture modeling via ML estimation, convergence is determined not only by the derivatives of the log-likelihood but also by the absolute and relative changes in the log-likelihood and the changes in the class counts (Muthén & Muthén, 2010). Convergence sometimes can be reached by increasing the number of iterations if the program stops before convergence due to the specified limited number of iterations or by using the preliminary parameter estimates as starting values. The model needs to be modified if

new starting values do not help. When an FMM achieves convergence, the solutions may still yield a local maxima. Therefore, it is important to use multiple sets of starting values during maximization to avoid local maxima and increase the chance of obtaining the best solution with the largest log-likelihood in confirmatory FMMs (Vermunt & Magidson, 2005). In Mplus, when the model doesn't converge in ML, the program will give a message indicating the program didn't stop normally. By requesting Tech 8, one can check if the highest log-likelihood value is replicated in at least two final stage solutions and provide the evidence that the model is converged to the global maximum solutions.

According to McLachlan and Peel (2000), the frequent occurrence of nonconvergence or convergence to local maxima can be due to the unbounded likelihood function in ML estimation. Class-varying loadings often yield instability, which leads to nonconvergence or local solution in FMMs with categorical outcomes. This phenomenon has been observed in many simulation studies for FMMs using ML estimation (Mann, 2009). Gagné (2004) found that nonconvergence is more likely in less restrictive FMMs in which loadings and thresholds are allowed to vary across classes. Lubke and Muthén (2007) found that when factor loadings vary across classes with indistinct class separation in terms of means, the models tend to be unstable and convergence rates may be poor. Hou and Hancock (2010) generated data from an IRT perspective and found relatively higher nonconvergence rates when item discrimination parameters varying across classes. With real data, the class-invariant loadings are often sufficient and used in behavioral science and educational

measurement (Clark, 2010; Jiao et al., 2010a, 2010b; Muthén, 2006; Muthén & Asparouhov, 2006).

Nonconvergence is also an issue for FMMs via Bayesian inference. To identify nonconvergence, it is recommended in Mplus to check whether the Proportional Scale Reduction (PSR) factor is close enough to 1 for each parameter (Asparouhov & Muthén, 2010b; Gelman et al., 2004; Gelman & Rubin, 1992). A value of 1.1 or less for all parameters indicates that the model converges (Gelman et al., 2004). As mentioned above, the iterative process produces a Markov chain of values drawn from the posteriors which are referred to as a *chain*. To judge convergence, the PSR compares the parameter variation within a chain to that across chains to make sure that the different chains do not converge to different values. Because it is possible that the PSR value may go up after a number of iterations, in order to gain further evidence of convergence, a longer chain is recommended (Asparouhov & Muthén, 2010a), especially for complex models like FMMs.

2.4.3.3 Label Switching

Arbitrary switching or assignment of class labels in mixture modeling is a wellknown issue and big problem in simulation study because the estimates cannot be meaningfully aggregated. There are several strategies to avoid the label switching issue. We followed the strategy for mixture modeling suggested by Asparouhov and Muthén (2010a) by introducing parameter constraints (inequalities) among the model parameters to constrain the classes, which does not mean the model is constrained but merely prevents the flipping of the labels across classes. Label switching often

happens when the classes are less distinguished, for example, if the classes are not well separated in terms of the mean, or if the parameters of the models of the classes are not so different. Also, it is more likely to happen when the sample size is small. In summary, label switching tends to happen when the sample size is small and/or the model parameters are not well distinguished across classes. Nylund et al. (2007) avoided label switching in their simulation study by specifying starting values equal to parameter generating values. This should be an alternative strategy for label switching in simulation studies.

2.5 Software Packages

There are three software packages that can handle factor mixture modeling or mixture IRT modeling with categorical outcomes. Mplus 6 (Muthén & Muthén, 2010) is the only one that can handle FMMs via both ML and Bayesian estimation methods. Both probit and logit links are available in ML estimation. Only the probit link is available in Mplus 6.1 when fitting FMMs via Bayesian estimation. The *mdltm* software (von Davier, 2005) can conduct mixture 1 Parameter Logistic (1PL) and 2 Parameter Logistic (2PL) IRT modeling via marginal ML estimation. And WinBUGS software (Spiegelhalter, et al, 2000) can handle mixture 1PL, 2PL and 3 Parameter Logistic IRT modeling via Bayesian inference. WINMIRA 2001 (von Davier, 2001) can be used to fit mixed Rasch models through conditional ML estimation.

Chapter 3: Method

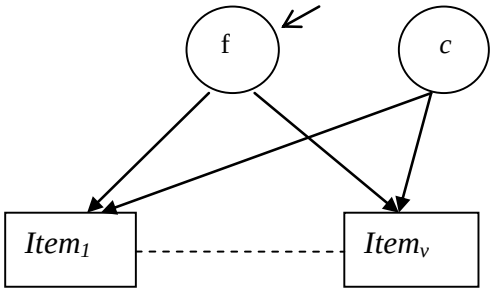
The main purpose of the simulation in the dissertation is to compare the two estimation methods (ML and Bayesian) in fitting the FMMs. This is investigated by comparing the performance of FMMs through the two estimation approaches in a Monte Carlo simulation where the sample data with known population parameters are generated. In this chapter, the descriptions of both data generating and data analysis models are provided along with the explanations of the manipulated and constant factors in the simulation study, followed by descriptions of varied simulation conditions and the criteria for evaluating the results. Because the goal of the study is not to explore the structure of the data, confirmatory analyses are conducted when we analyze the generated data. The number of factors and latent classes, the relation between the factors and items, and whether model parameters are class-specific are considered as the prior knowledge before the analyses. All the datasets are generated and analyzed in Mplus 6.1 (Muthén & Muthén, 2010) and SAS 9.0.

3.1 Data Generating Models

Figure 3-1 illustrates the general FMMs considered in the data generating in this study. The letter *c* stands for the categorical latent class variable with *K* classes. The letter *f* represents the continuous latent factor. The binary indicators are represented by the boxes $Item_1$ to $Item_v$, where *v* is the number of items with binary outcomes. The arrows from the latent class variable *c* only point to the items indicating that the distinction of latent groups only depends on the perspective of item thresholds or difficulties. The arrows from the factor to the items indicate that the

factor is indicated by the these binary variables. The arrow pointing to the factor *f* indicates the variance of the factor may be modeled. In summary, the factor loadings are class-invariant whereas the item thresholds are noninvariant across classes in the data generating models. In practice, the class-invariant loadings are widely used in behavioral science and educational measurement (Clark, 2010; Jiao, et al., 2010a, 2010b; Muthén, 2006; Muthén & Asparouhov, 2006). The factor means and covariance matrices can be class-invariant or noninvariant, depending on the previous knowledge or theory. However, when distributions of latent factors are different across classes, the difference between discrepancies in item difficulties across latent classes might be confounded with the difference between the latent factors (Kelderman & Macready, 1990). In the current study, the distribution of the factor is not varying across classes in the generated models.

Figure 3-2. FMMs in the simulation.



Binary outcomes are generated in the study, which is quite common in the behavioral sciences and measurement. Binary outcomes provide limited information in estimation compared with continuous variables (Asparouhov & Muthén, 2010b). Unknown class membership makes the estimation more difficult. For each simulation condition, 25 replications were generated, given the large amount of time for

estimating FMM with binary outcomes, especially when 50 random starts values are required for each replication. The similar number of replications and random starts are often found in other simulations on mixture modeling, especially with categorical outcomes that reasonably well served the purposes of those simulations (Dai, 2009; Jiao, et al., 2010a, 2010b, Li et al., 2009; Lubke, 2006).

Some conditions in the simulation are kept constant, given that it is not feasible to investigate all possibilities in a reasonable time frame. In the current study, 16 binary datasets were generated under different unidimensional factor mixture models with two latent classes of equal class proportions. Then 16 confirmatory data analysis models are fitted via different estimation methods (i.e., ML and Bayesian with 3 sets of priors), resulting in 64 investigated conditions in total.

The latent traits of the individuals are normally distributed with a mean of zero and variance of one in both latent classes, because the current study is designed to investigate the performance of ML and Bayesian methods in recovering the factor loadings and threshold parameters across classes. Without separation in mean structure across classes, the difference between discrepancies in item parameters across latent classes would not confound with the difference between the latent factors. This study can provide useful information for the studies such as DIF, measurement invariance and detecting latent groups who use different strategies to solve test items in the presence of an invariant mean structure of latent factors. In addition, this specification tends to make the convergence in the model estimation more difficult than one in which the means of latent groups are distinctly separated (Dai, 2009). In other words, it is expected that the convergence in the estimation

would take less time with higher convergence rate if there is larger mean separation as well as distinct item parameters across classes than the simulation conditions in the current study.

3.1.1 Manipulated Factors in the Simulation Study

The most important factor manipulated in the study is the estimation method. Both ML and Bayesian inference are employed to fit the FMMs. In Bayes runs, we use three different normally distributed, weakly informative priors for loading and threshold parameters in fitting FMMs given that the model estimates are on the probit scale. Normally distributed priors are widely used for loadings and threshold parameters in the literature, and they do a reasonable job in providing the accurate and stable estimates. Noninformative priors are conventionally employed in the vast of literature on Bayesian inference. However, weakly informative priors are recommended and used by Gelman et al. (2008a) and Asparouhov and Muthén (2010b) for logistic or probit regression with binary outcomes. For detailed reviews of the priors used for Bayesian estimation, see the previous chapter. The threshold parameters are usually within the range of $\pm$ 2, which correspond to the item difficulty parameters of range of $\pm$ 2.5 given the investigated factor loadings of 0.8 and 0.4 in the current study. The loading of 0.8 is a relative high loading for FMMs with categorical outcomes. In the current study, the priors of N(0,10^5), N(0, 5), and N (0, 1) are chosen to represent noninformative and weakly informative priors for the model parameters. Table 3-1 shows the priors' variance and their 90% and 95% limits.

Table 3-1. Priors for loadings and thresholds and corresponding 90% & 95% limits

Variance	90% limits	95% limits
1	$\pm$1.64	$\pm$1.96
5	$\pm$3.68	$\pm$4.38
10^5	$\pm\infty$	$\pm\infty$

Another manipulated factor is the number of indicators within each class. Some researchers believe that with a larger number of indicators, more information can be used in the estimation procedure, resulting in less biased estimates. For example, Marsh et al. (1998) and Gagné and Hancock (2006) found that parameter estimates were less biased via ML estimation as the number of indicators per factor increased with sample size held constant. Li et al. (2009) found that the estimates were less biased as test length increased for mixture 2PL IRT models via Bayesian estimation. On the other hand, Asparouhov and Muthén (2010b) pointed out that having a large number of indicators is more challenging than one would expect in MCMC and good estimates need a large sample size when the number of indicators is large (Asparouhov & Muthén, 2010b). The values such as 8, 10, 15, 30, and 36 for the number of indicators per factor have been applied in various FMMs and mixture IRT simulation studies (Dai, 2009; Jiao & von Davior, 2010; Li et al., 2009; Mann, 2009; Nylund et al., 2007). In this study, we evaluate the impact of the number of indicators on the estimates by simulating a medium number of indicators (the number of items is 30) and a small number of indicators (the number of items is 8) for the one factor in testing context.

The third manipulated factor is the sample size. Four different sample sizes are chosen in the study in the hope of finding the trend of the changes in estimation accuracy as the sample size increases. As seen in the last chapter, in the Bayesian framework various noninformative priors are used for the loadings and threshold; however, in most studies the sample sizes were large and the effect of the priors on estimation was very small. Relatively small sample sizes were not studied. Muthén and Asparouhov (2010) explored the conditions with a small sample size of 50 to 500 for factor analysis modeling with binary outcomes in Bayesian analysis. Sample sizes of 200, 500, and 1000 were used in a Monte Carlo simulation for a similar FMM with binary data in ML estimation (Nylund et al., 2007). Sample sizes of 1,000 and 5,000 were used in mixture 1PL mixture model in ML and Bayesian runs, respectively (Dai, 2009; Jiao, 2010a, 2010b), and sample sizes of 600 and 1,200 were used in Li's paper (Li et al., 2009) for 2-PL mixture IRT model via Bayesian estimation. In this study, a pilot study showed that a sample size of 500 is sufficient for successful convergence when the number of indicators is 8. Therefore, we use a sample size of 500 as the starting point, and sample sizes of 1,000, 2,000, and 5,000 are also investigated. We also hope that at least a range of sample sizes can be found at which there is not much difference between ML and Bayesian estimation performance in fitting FMMs.

The impact of the magnitude of loadings on parameter recovery and classification is also investigated in the current study. Previous research showed that increases in loading magnitude usually tend to yield less biased model parameter estimates for confirmatory FA models with ML estimation (Gagné & Hancock, 2006). In the current study, the values of 0.8 and 0.4 are used to represent different

magnitudes of loadings in the factor models, which correspond to the values of 0.8 and 0.4 item discrimination parameters in mixture 2PL IRT models when residual variances are fixed at 1 by Mplus. The threshold parameters in FMMs are chosen at the values of ± 0.5. Recall the Equation (2-22), thus the corresponding values of the item difficulty parameters are $\pm$ 1.25 and $\pm$ 0.63 in mixture 2PL IRT models.

Table 3-2. Simulation conditions

Condition	Sample size	Estimation method	# of items	Magnitude of loadings
1	500	ML	8	.8
2	500	Bayes prior ~N(0,1)	8	.8
3	500	Bayes prior~N(0,5)	8	.8
4	500	Bayes prior~N(0,10^5)	8	.8
5	1000	ML	8	.8
6	1000	Bayes prior~N(0,1)	8	.8
7	1000	Bayes prior~N(0,5)	8	.8
8	1000	Bayes prior~N(0, 10^5)	8	.8
9	2000	ML	8	.8
10	2000	Bayes prior~N(0,1)	8	.8
11	2000	Bayes prior~N(0,5)	8	.8
12	2000	Bayes prior~N(0, 10^5)	8	.8
13	5000	ML	8	.8
14	5000	Bayes prior~N(0,1)	8	.8
15	5000	Bayes prior~N(0,5)	8	.8
16	5000	Bayes prior~N(0, 10^5)	8	.8
17	500	ML	8	.4
18	500	Bayes prior ~N(0,1)	8	.4
19	500	Bayes prior~N(0,5)	8	.4
20	500	Bayes prior~N(0, 10^5)	8	.4
21	1000	ML	8	.4
22	1000	Bayes prior~N(0,1)	8	.4
23	1000	Bayes prior~N(0,5)	8	.4
24	1000	Bayes prior~N(0, 10^5)	8	.4
25	2000	ML	8	.4
26	2000	Bayes prior~N(0,1)	8	.4
27	2000	Bayes prior~N(0,5)	8	.4
28	2000	Bayes prior~N(0, 10^5)	8	.4
29	5000	ML	8	.4
30	5000	Bayes prior~N(0,1)	8	.4
31	5000	Bayes prior~N(0,5)	8	.4

32	5000	Bayes prior~N(0, 10^5)	8	.4
33	500	ML	30	.8
34	500	Bayes prior~N(0,1)	30	.8
35	500	Bayes prior~N(0,5)	30	.8
36	500	Bayes prior~N(0, 10^5)	30	.8
37	1000	ML	30	.8
38	1000	Bayes prior~N(0,1)	30	.8
39	1000	Bayes prior~N(0,5)	30	.8
40	1000	Bayes prior~N(0, 10^5)	30	.8
41	2000	ML	30	.8
42	2000	Bayes prior~N(0,1)	30	.8
43	2000	Bayes prior~N(0,5)	30	.8
44	2000	Bayes prior~N(0, 10^5)	30	.8
45	5000	ML	30	.8
46	5000	Bayes prior~N(0,1)	30	.8
47	5000	Bayes prior~N(0,5)	30	.8
48	5000	Bayes prior~N(0, 10^5)	30	.8
49	500	ML	30	.4
50	500	Bayes prior~N(0,1)	30	.4
51	500	Bayes prior~N(0,5)	30	.4
52	500	Bayes prior~N(0, 10^5)	30	.4
53	1000	ML	30	.4
54	1000	Bayes prior~N(0,1)	30	.4
55	1000	Bayes prior~N(0,5)	30	.4
56	1000	Bayes prior~N(0, 10^5)	30	.4
57	2000	ML	30	.4
58	2000	Bayes prior~N(0,1)	30	.4
59	2000	Bayes prior~N(0,5)	30	.4
60	2000	Bayes prior~N(0, 10^5)	30	.4
61	5000	ML	30	.4
62	5000	Bayes prior~N(0,1)	30	.4
63	5000	Bayes prior~N(0,5)	30	.4
64	5000	Bayes prior~N(0, 10^5)	30	.4

The five manipulated factors result in a total of 64 simulation conditions. Table 3-2 shows each combination of the manipulated factors. Table 3-3 summarizes the constant and manipulated factors in this study. Tables 3-4 to 3-7 are the model specifications for data generating when the number of items are eight and thirty across all the sample sizes.

Table 3-3. Summary of Monte Carlo Population Specifications for FMMs with Binary outcomes

Population specifications	
Sample size	500, 1000, 2000, 5000
Replication	25
Number of items	8, 30
Population number of classes	2
Class proportion	50%:50%
Estimation	ML, Bayes
Priors for Bayes	Loading and threshold : 3 different priors

Table 3-4. FMM data generating specification with 8 items and loading of 0.8

Model parameters	Class 1		Class 2	
Class size	50%		50%	
Factor variance	1		1	
Factor means	0		0	
Item	τ (b)	λ (a)	τ (b)	λ (a)
1	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
2	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
3	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
4	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
5	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
6	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
7	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
8	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)

Table 3-5. FMM data generating specification with 8 items and loading of 0.4

Model parameters	Class 1		Class 2	
Class size	50%		50%	
Factor variance	1		1	
Factor means	0		0	
item	τ (b)	λ (a)	τ (b)	λ (a)
1	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
2	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
3	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
4	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
5	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
6	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
7	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
8	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)

Table 3-6. FMM data generating specification with 30 items and loading of 0.8

Model parameters	Class 1		Class 2	
Class size	50%		50%	
Factor variance	1		1	
Factor means	0		0	
Item	τ (b)	λ (a)	τ (b)	λ (a)
1	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
2	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
3	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
4	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
5	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
6	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
7	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
8	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
9	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
10	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
11	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
12	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
13	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
14	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
15	.5 (.63)	.8 (.8)	-.5 (-.63)	.8 (.8)
16	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
17	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
18	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
19	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
20	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
21	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
22	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
23	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
24	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
25	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
26	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
27	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
28	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
29	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)
30	-.5 (-.63)	.8 (.8)	.5 (.63)	.8 (.8)

Table 3-7. FMM data generating specification with 8 items and loading of 0.4.

Model parameters	Class 1		Class 2	
Class size	50%		50%	
Factor variance	1		1	
Factor means	0		0	
Item	τ (b)	λ (a)	τ (b)	λ (a)
1	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
2	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
3	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
4	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
5	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
6	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
7	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
8	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
9	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
10	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
11	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
12	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
13	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
14	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
15	.5 (1.25)	.4 (.4)	-.5 (-1.25)	.4 (.4)
16	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
17	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
18	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
19	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
20	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
21	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
22	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
23	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
24	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
25	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
26	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
27	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
28	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
29	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)
30	-.5 (-1.25)	.4 (.4)	.5 (1.25)	.4 (.4)

3.2 Data Analysis Models

Starting values which are equal to population values are also given to the estimated parameters as the basis for the perturbation tht is done in estimation. Given that the outcome variable is binary, which makes estimation relatively time-consuming, fifty sets of random starts are applied with ten iterations for each set in ML estimation as the basis for the perturbation that is done in estimation, consistent with previous research (Lubke & Muthén, 2007; Lubke & Neale, 2006, 2008; Mann, 2009).

A total of 2,500 burn-in iterations are suggested for the 1PL and 2PL mixture model conditions by Gelman and Rubin (1992). Because it is observed that the PSR value may go up after *q* iterations (Muthén & Asparouhov, 2010) before achieving the stable PSR value, in the Bayesian runs of the current study 50,000 iterations with the first half as a burn-in phase are conservatively required to ensure real convergence in pilot study for 8 conditions with 30 items, though a much lower number of iterations were found in many previous studies (Dai, 2009; Gelman & Rubin, 1992; Li et al., 2009). It was found that 10,000 iterations is sufficient for model convergence in the pilot study, therefore, 10,000 iterations was finally used for all the conditions in the main study.

A confirmatory structure is assumed in the study, in that the number of factors, the relation between the factor and items, and whether model parameters are class-specific or class-invariant are known before the analyses. In the analysis models, factor means are constrained to be equal across classes and set to zero for giving the origin of the factor (Clark, 2010; Mann, 2009; Muthén, 2008; Yung, 1997),

which means there is no mean separation across two classes in a real scenario. In addition, the factor variance is fixed to the true value of one for each class to provide a metric for the factor. All thresholds and loadings are free to be estimated except the constraint of equal loadings across classes.

3.3 Evaluation of Estimation Outcomes

The study compares the two estimation methods not only by providing theoretical backgrounds of the two estimation methods but more importantly by evaluating estimation outcomes that include the analyses of parameter recovery and classification accuracy.

3.3.1 Parameter Recovery

Bias, relative bias, and standard error (*SE*) are used in the recovery analysis for comparison.

The bias of an estimator, as a measure of the accuracy of the parameter estimate, is the difference between an estimator's expected value and the true value of the parameter being estimated. For ML estimates, it can be expressed as

$$BIAS_{ML} = E\left[\left(\hat{\theta}_{ML} - \theta_{true}\right)\right]. \quad (3\text{-}1)$$

And for Bayesian estimates,

$$BIAS_{Bayes} = E[(\hat{\theta}_{Bayes} - \theta_{true})]. \quad (3\text{-}2)$$

The relative bias of an estimator is the value of bias divided by the true value of the parameter, and can be calculated as

$$RELATIVE\ BIAS_{ML} = \frac{BIAS_{ML}}{\theta_{true}}. \quad (3\text{-}3)$$

And for Bayesian estimates

$$RELATIVE\ BIAS_{Bayes} = \frac{BIAS_{Bayes}}{\theta_{true}}.$$

(3-4)

The standard error (*SE*) is a measure of the stability of estimation. Suppose *n* is the number of replications, *SE* for the ML estimator for model parameter θ for each item is defined as

$$SE_{ML} = \sqrt{\frac{1}{n-1}\sum_1^n(\hat{\theta}_{ML} - \overline{\hat{\theta}}_{ML})^2}\ . \quad (3\text{-}5)$$

And for the Bayes estimator, *SE* for each item is defined as

$$SE_{Bayes} = \sqrt{\frac{1}{n-1}\sum_1^n(\hat{\theta}_{Bayes} - \overline{\hat{\theta}}_{Bayes})^2}\ . \quad (3\text{-}6)$$

3.3.2 Classification Accuracy

The estimated class memberships based on posterior class probabilities are compared to the population values. Classification accuracy is investigated in terms of both the average percentage of correct classification in each condition and scatterplots of posterior class probabilities in select conditions.

Chapter 4: Results

The study compares the estimation methods not only by providing theoretical backgrounds of the two methods but more importantly by evaluating estimation outcomes that include the analyses of parameter recovery and classification accuracy.

4.1 Parameter Recovery

Bias, relative bias, and standard error (*SE*) are used in the recovery analysis for the comparison purposes. Bias and relative bias reflect the difference between the estimated and the true value of the parameters. *SE* measures the stability of the estimation. Appendix A Tables A-1 to A-3 show the averages of biases, relative biases and *SE*s of the model parameters estimates over the items in the total of 64 investigated conditions.

Figures 4-1 to 4-6 reflect the difference of the averages of biases and *SE*s over items among the estimation methods given different combination of three manipulated factors (i.e. the number of items, magnitude of loadings and sample size). These combinations are represented by 16 numbers in the horizontal axis in Figures 4-1 to 4-6, which are divided into 4 cohorts of sample sizes in Table 4-1. In each cohort the sample size orders from 500, 1000, 2000 to 5000. In these 6 figures, Bayes1 represents the Bayesian estimation with weakly informative prior of N(0, 1), Bayes2 the Bayesian estimation with weakly informative prior of N(0, 5), and Bayes3 the Bayesian estimation with noninformative prior of N(0, 10^5).

Table 4-1. Combination of manipulated factors

Cohort	number	combination
1	1~4	8 items & 0.8 loading
2	5~8	8 items & 0.4 loading
3	9~12	30 items & 0.8 loading
4	13~16	30 items & 0.4 loading

Figure 4-1. Bias of loading

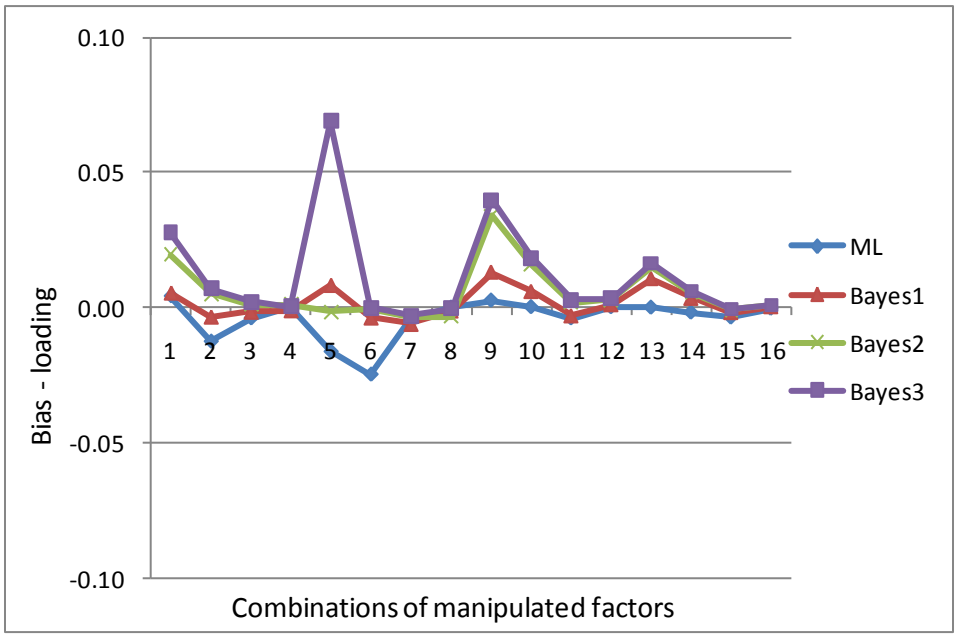


Figure 4-2. Bias of class 1 threshold

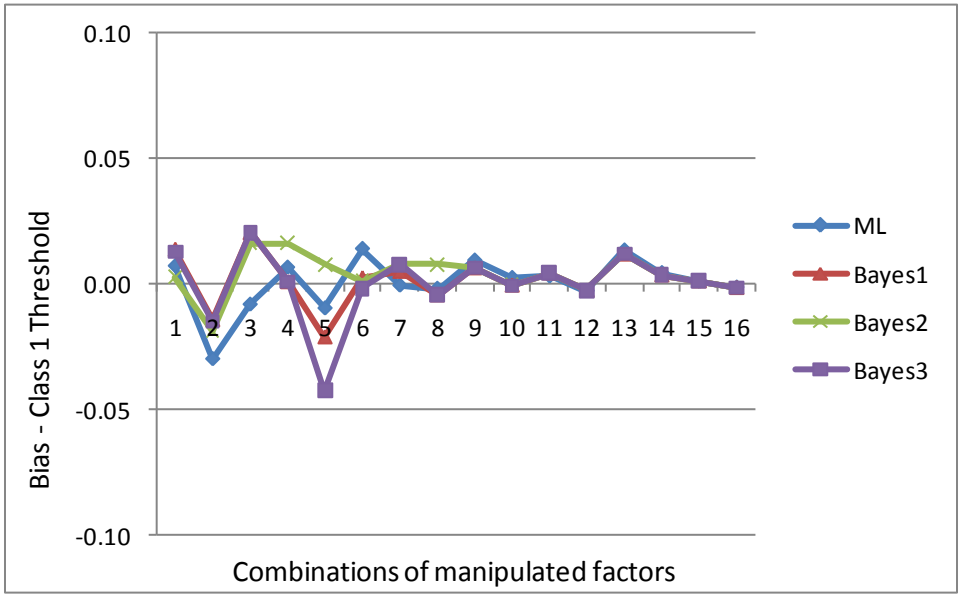
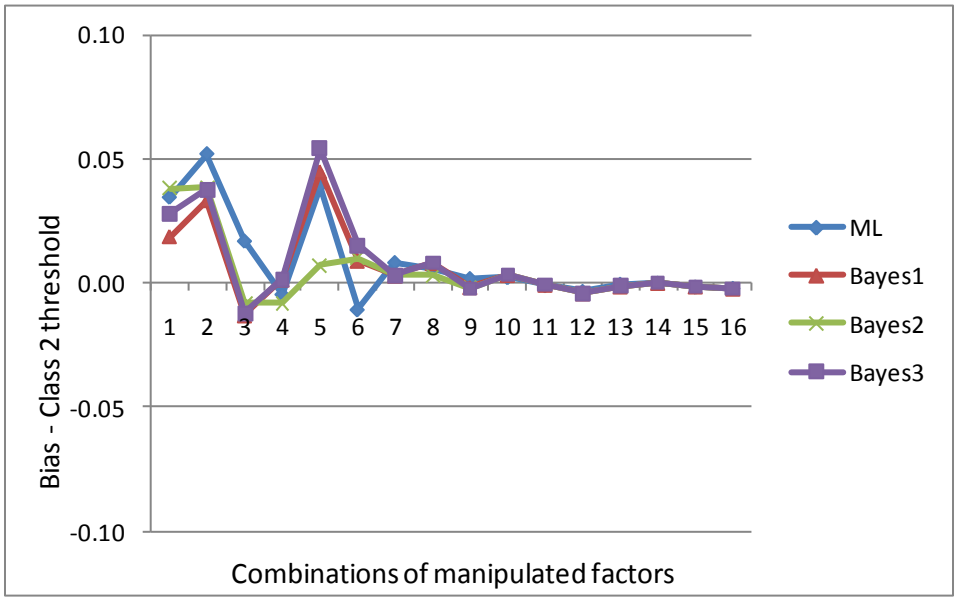


Figure 4-3. Bias of class 2 threshold



Based on Figures 4-1 to 4-3, the general findings for the averages of biases of both loadings and thresholds are:

(1) When the sample size increases, the bias decreases. The bias is large when sample size is small for each cohort. It is especially the case when the number of item is 8.

(2) When number of items increases, the bias tends to decrease. The bias is large when the number of items is 8. It is especially the case when the sample size is 500 and 1000.

(3) Estimation methods make differences in term of the bias for most conditions. Bayesian with noninformative priors generally yield larger bias than other estimation methods especially when the loading is 0.4 and the number of items is 8. When number of items is 30, the biases from different estimation methods are very close for threshold estimates.

(4) Loading bias is more influenced by estimation methods than threshold bias is, especially for the conditions with 8 items.

Figure 4-4. SE of loading

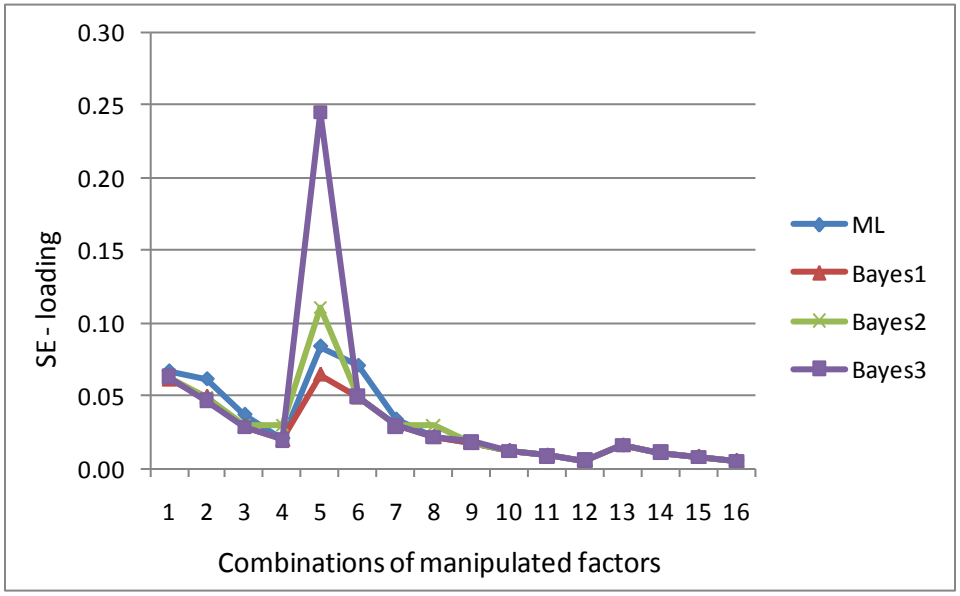


Figure 4-5. SE of class 1 threshold

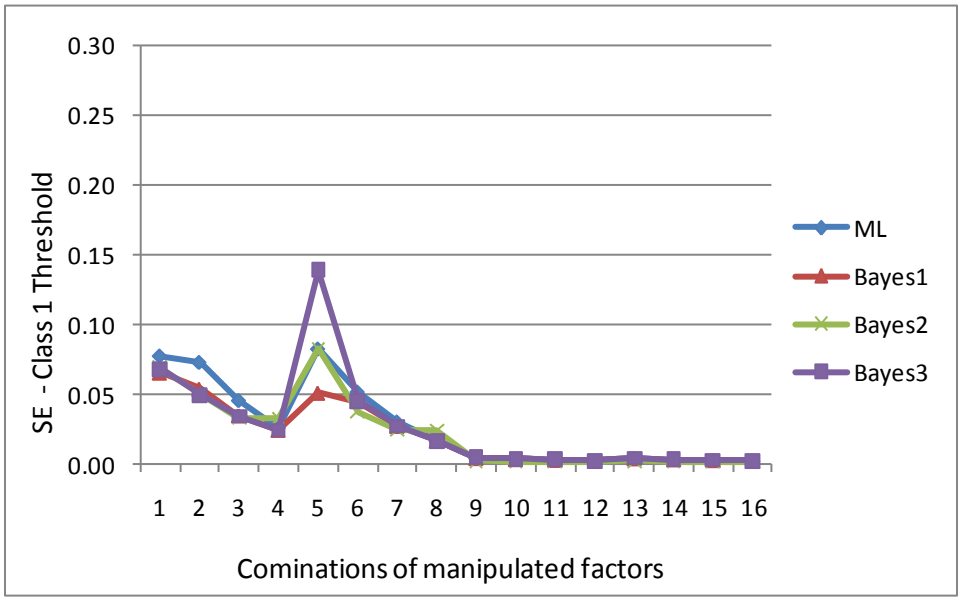
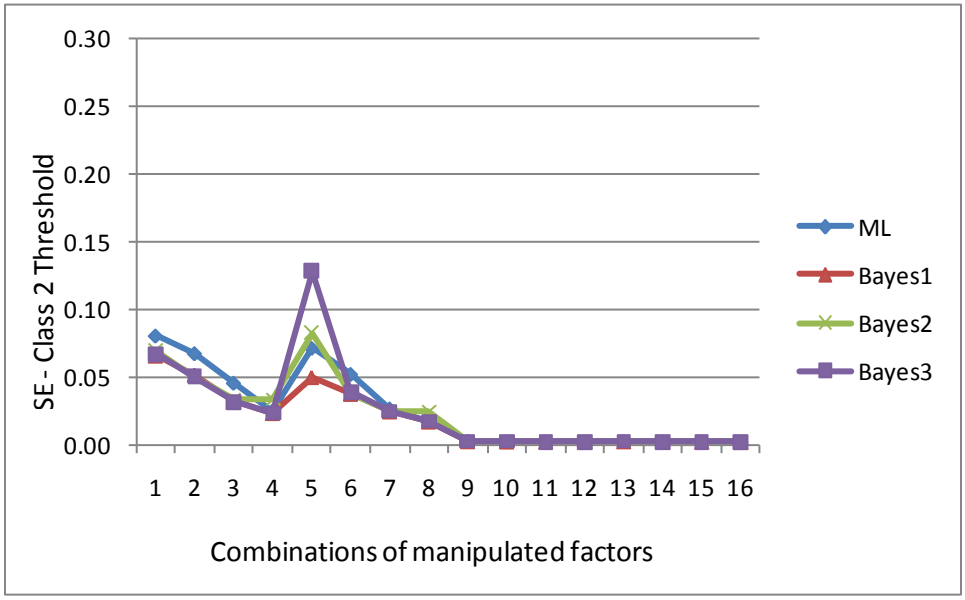


Figure 4-6. SE of class 2 threshold



Based on Figures 4-4 to 4-6, the general findings for the *SE* of loadings and thresholds over items are:

(1) When sample size increases, the *SE* decreases. The *SE* is large when sample size is small for each cohort. Threshold *SE* is less influenced by sample size when the number of items is 30 than the ones with 8 items.

(2) When number of items increases, the *SE* decreases. The *SE* is large when the number of items is 8. This is especially true when the sample size is 500 and 1000.

(3) Estimation methods make a difference in terms of *SE*s when the number of items is 8. Bayesian with noninformative priors generally yield bigger *SEs* than other estimation methods when the loading is 0.4. ML yields larger *SEs* than other estimation methods when the loading is 0.8. When the

number of items is 30, the *SE*s from the different estimation methods are very close.

(4) Loading *SEs* are more influenced by estimation methods than threshold *SEs* in Bayesian estimation, especially for the conditions with 8 items. A noninformative prior tends to produce the largest *SE* and a prior of N(0,1) the smallest *SE* in the cohort 2.

In the following sections, the further comparisons among the estimation methods in terms of relative biases of loading and threshold estimates are discussed.

4.1.1 Recovery of Loading Parameters

Figures 4-7 to 4-10 give some examples of visual illustrations on how well the parameters are recovered by ML and Bayesian estimation methods in 4 different manipulated conditions. For the graphs of all other investigated conditions, see Appendix B.

In all these figures, the solid blue curve represents the estimates from ML methods; the solid green curve represents the calibration from Bayesian methods with weakly priors of N(0, 1); the solid purple curve stands for the estimates from Bayesian methods with priors of N(0,5), and dotted red curve are the estimates from Bayesian with noninformative priors. The horizontal axis is item number, and the vertical axis is relative bias. Figures B-1 to B-8 depict the results from the conditions in which the number of items is 8, and Figures 4-7 to 4-10 and B-25 to B-28 display the results from the conditions in which the number of items is 30.

When sample size is relatively small (e.g., 500) and the number of item is 30, ML generally tends to yield less biased loading estimates than Bayesian estimation. As the sample size increases, the difference of relative bias between ML and Bayesian decreases; when the sample size is 5000, the differences in relative bias are hardly inspected visually. Among the Bayesian methods, the priors of N (0, 1) yield estimates with smaller relative biases than the priors of N(0,5) do, and the priors of N(0,5) yield smaller relative biases than the noninformative priors do. In addition, the larger the sample sizes are, the less biased the loading estimates are for both ML and Bayesian methods.

When the number of items is 8, the performance differences among estimation methods are not as obvious as the ones with 30 items, which are shown in the Appendix B Figures B1 to B8. However, it is clear that as the sample size increases, the relative biases of loading parameters decrease with all the estimation methods.

The magnitude of loadings affected the parameter estimates of loadings in two ways. First, in ML estimation, the sign of loading estimates flipped frequently with a true value of 0.4, which is not a problem for the conditions with a true value of 0.8. In other words, if the values of loading parameters are around or below medium such as 0.4 in the current study, the flipping of the loading signs is very likely in ML. If one ignores the occurrence of it in simulations, ML bias for loading parameters would be inflated due to their sign flipping. However, in Bayesian estimation, the flipping never happened in the investigated conditions. Second, the magnitude of loadings appears to impact the relative biases of estimates for all the estimation methods with small sample size, though it seems not to make any estimation method superior. For

the conditions with 8 items, when sample size is as small as 500, the relative biases in the conditions with low magnitude of loadings tend to be larger than those with high magnitude of loadings (e.g., Appendix Figure B1 vs. B5). For the conditions with 30 items, high magnitude of loadings tends to yield less biased estimates than low magnitude of loadings (e.g. Figure 4-9 vs. B-27).

Figure 4-7. Relative bias of loading for each item when N = 500, loading = 0.8 and number of items = 30

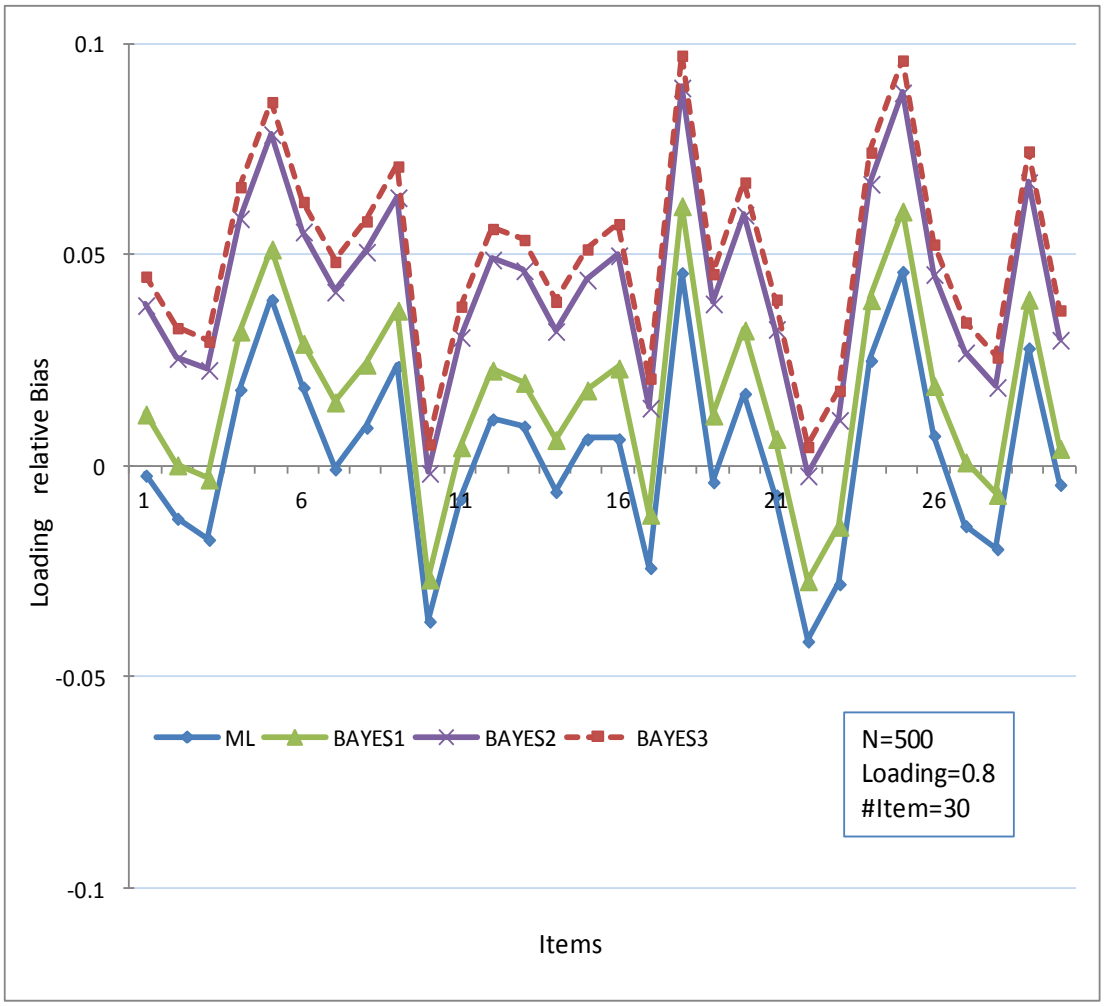


Figure 4-8. Relative bias of loading for each item when N = 1000, loading = 0.8 and number of items = 30

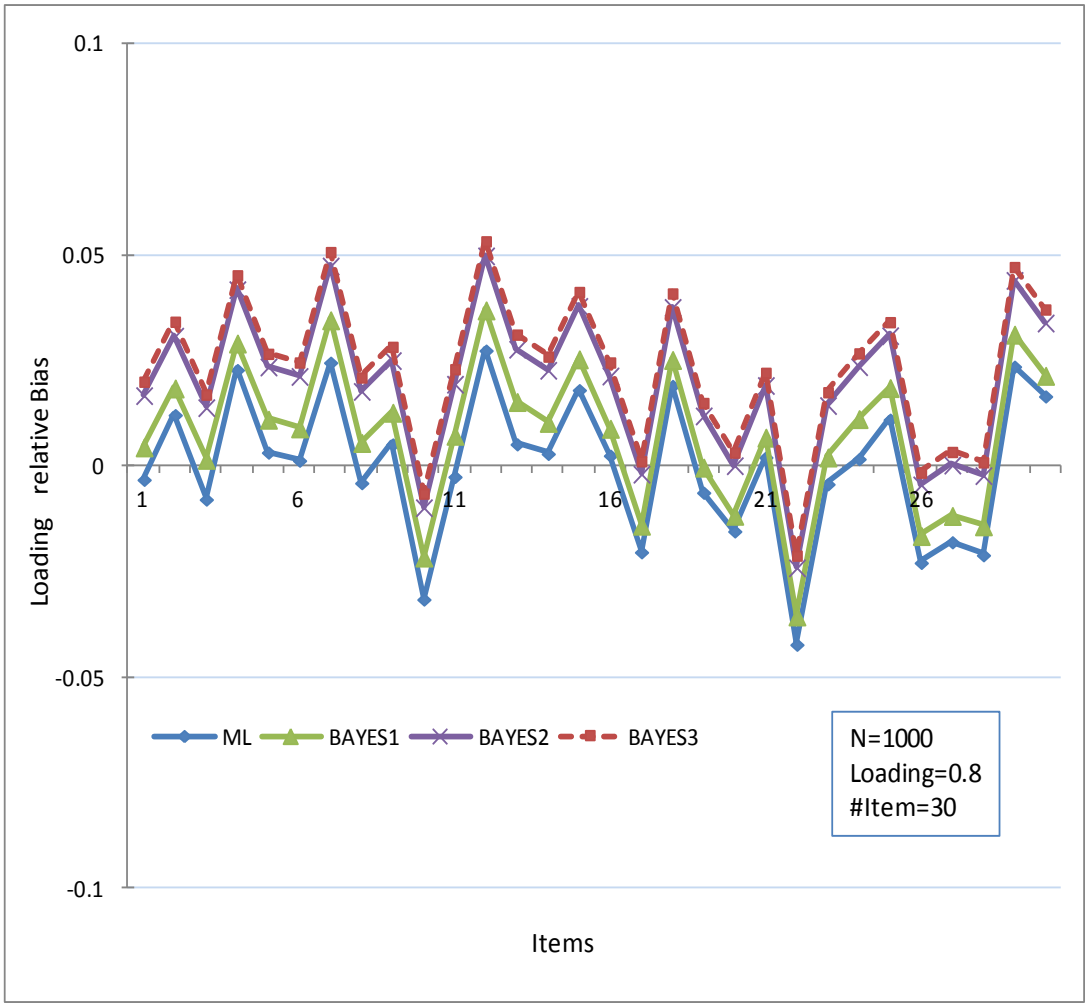


Figure 4-9. Relative bias of loading for each item when N = 2000, loading = 0.8 and number of items = 30

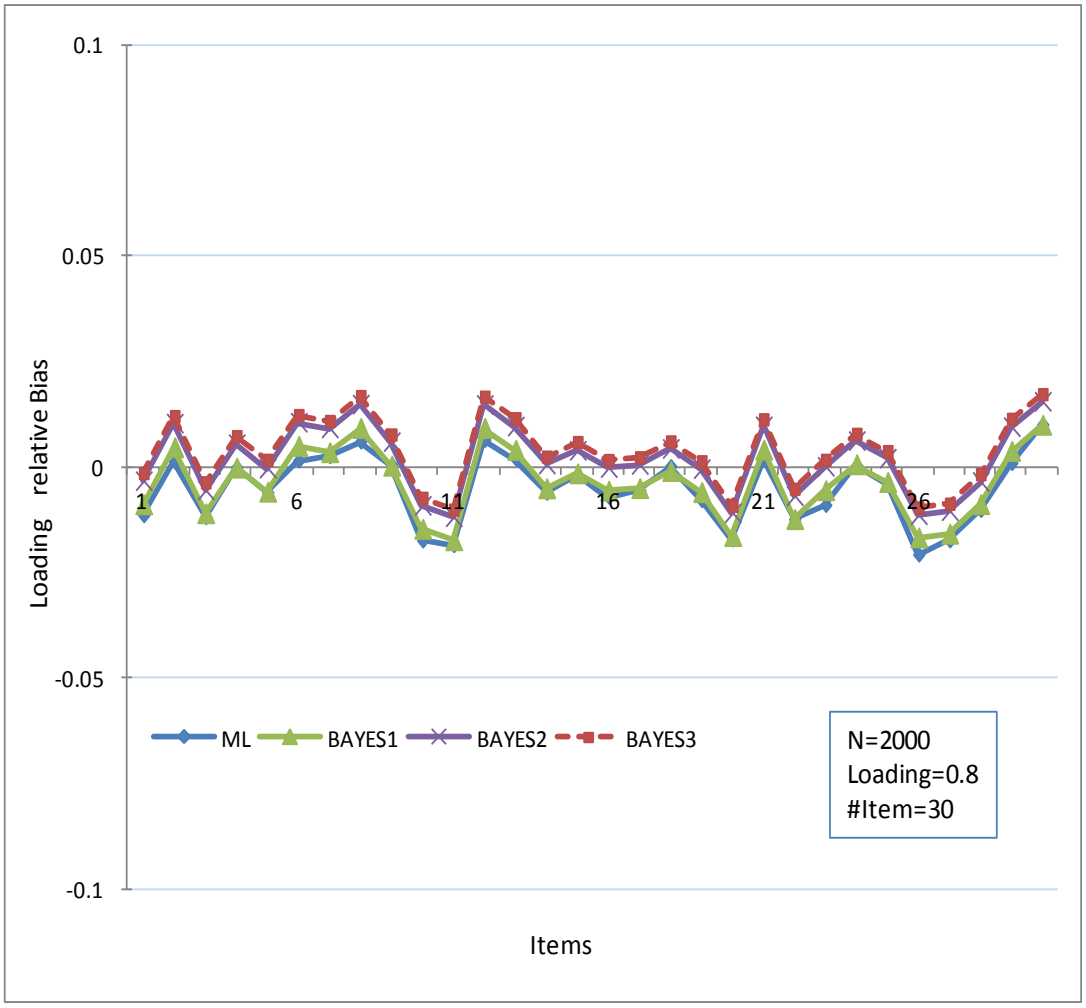
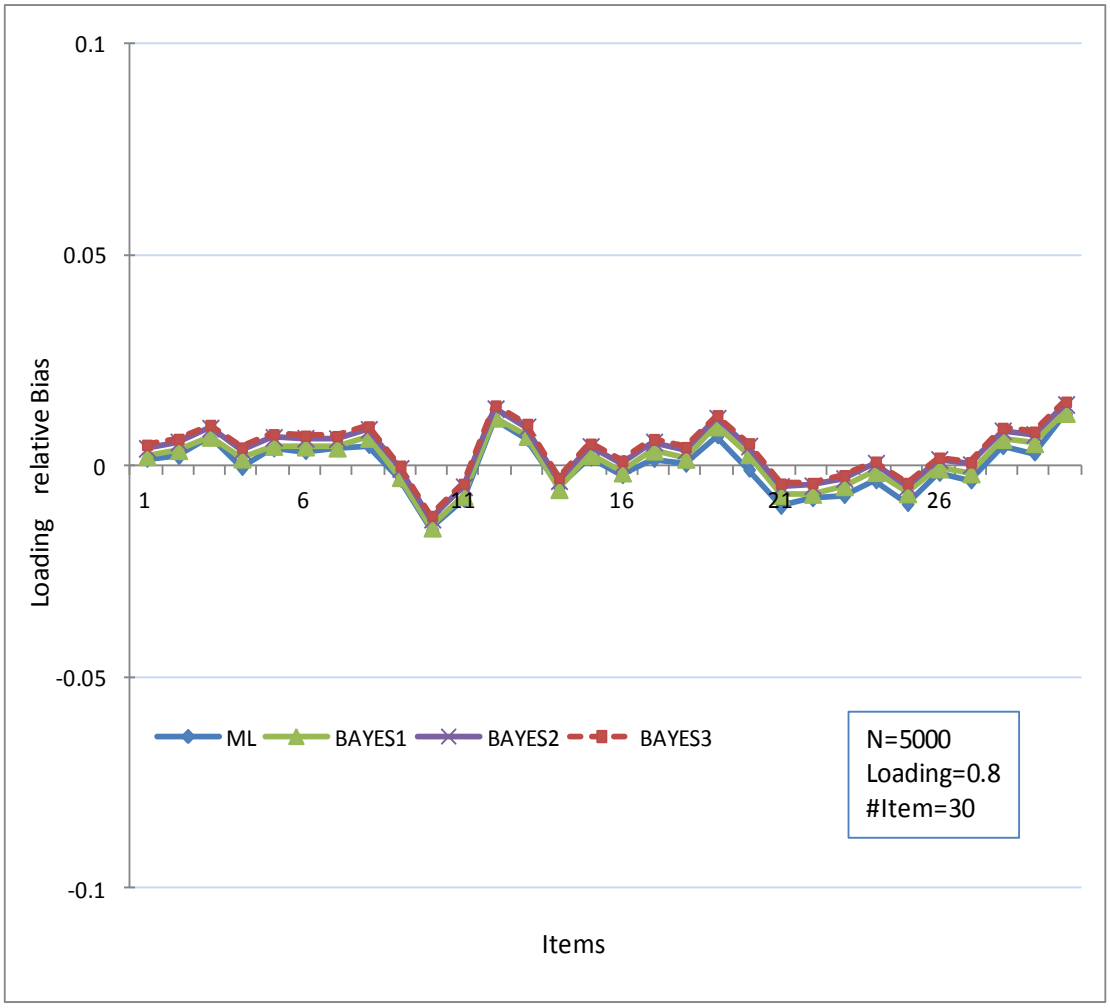


Figure 4-10. Relative bias of loading for each item when N=5000, loading=0.8 and number of items = 30



Besides visual inspection for the performance difference among the estimation methods, four-way ANOVA (sample size * number of items * magnitude of loading * estimation method) on the averaged relative bias of loading parameters over items was conducted. All main and interaction effects were included in the following analyses and the follow-up tests were focused on the significant effects related to the estimation methods. Three significant effects were found: a significant main effect of estimation methods, F(3, 1492) = 7.039, p < 0.01, η_p^2= 0.014; a significant main effect of sample sizes, F(3, 1492) = 10.722, p< 0.01, η_p^2= 0.021; and a significant interaction of estimation methods by sample sizes, F(9, 1492) = 2.675, p < 0.01, η_p^2= 0.016. In order to compare the different estimation methods, four one-way ANOVAs on the average relative bias across items at each level of sample size were conducted and results can be found in Table 4-2.

Table 4-2 One-way ANOVAs for impact of estimation method on four levels of sample sizes on Relative Bias of Loadings

Sample size	*F*	*P*	η_p^2
500	*F*(3, 373) = 4.324	0.005	0.034
1000	*F*(3, 382) = 2.777	0.041	0.021
2000	*F*(3, 390) = 1.356	0.256	0.010
5000	*F*(3, 394) = 0.188	0.904	0.001

For sample size of 500, Tukey post-hoc comparison of the estimation methods indicates that the ML estimation method (M = -0.0083, 95% CI [-0.0321, 0.0154]) performed significantly better than the Bayesian method with noninformative priors (M = 0.0673, 95% CI [0.0187, 0.1159]), p = 0.002. The comparisons of other pairs of

estimation methods were not statistically significant at $p < 0.05$. For sample size of 1000, Tukey post-hoc comparisons among the estimation methods didn't show any significant result, though the ML method (M = -0.0207, 95% CI [-0.0522, 0.0108]) marginally non-significantly outperformed the Bayesian method with noninformative priors (M = 0.0119, 95% CI [0.0032, 0.0206]), $p = 0.052$.

4.1.2 Recovery of Class 1 Threshold Parameters

One can tell the performance difference among the different estimation methods from the following Figures 4-11 to 4-14 especially when sample size is 500 and 1000. However, compared with loading parameters, the difference between ML and Bayesian recovery of thresholds parameters are much less influenced by the change of sample sizes, though as the sample size increases the relative biases of estimates greatly decreases and very close to each other among all the estimation methods. For the details, see Figures 4-11 to 4-14, Appendix B Figures B9 to B16, and B29 to B32.

The relative biases are generally larger in the conditions with 8 items than the ones in the conditions with 30 items, especially when sample size is small such as 500 and 1000 (e.g., Figures 4-12 vs. Appendix B Figure B10). With 30 items, the performance difference among the estimation methods is very small.

It is also found in the graphs that the magnitude of loadings affected the parameter estimates in some conditions. When sample size and number of item are small, say, 500 and 8 respectively, the relative biases in the conditions with low

magnitude of loadings tend to be larger than those with high magnitude of loadings (e.g., Appendix B Figures B9 vs. B13).

Figure 4-11. Relative bias of class 1 threshold for each item when N = 500, loading = 0.8 and number of items = 30

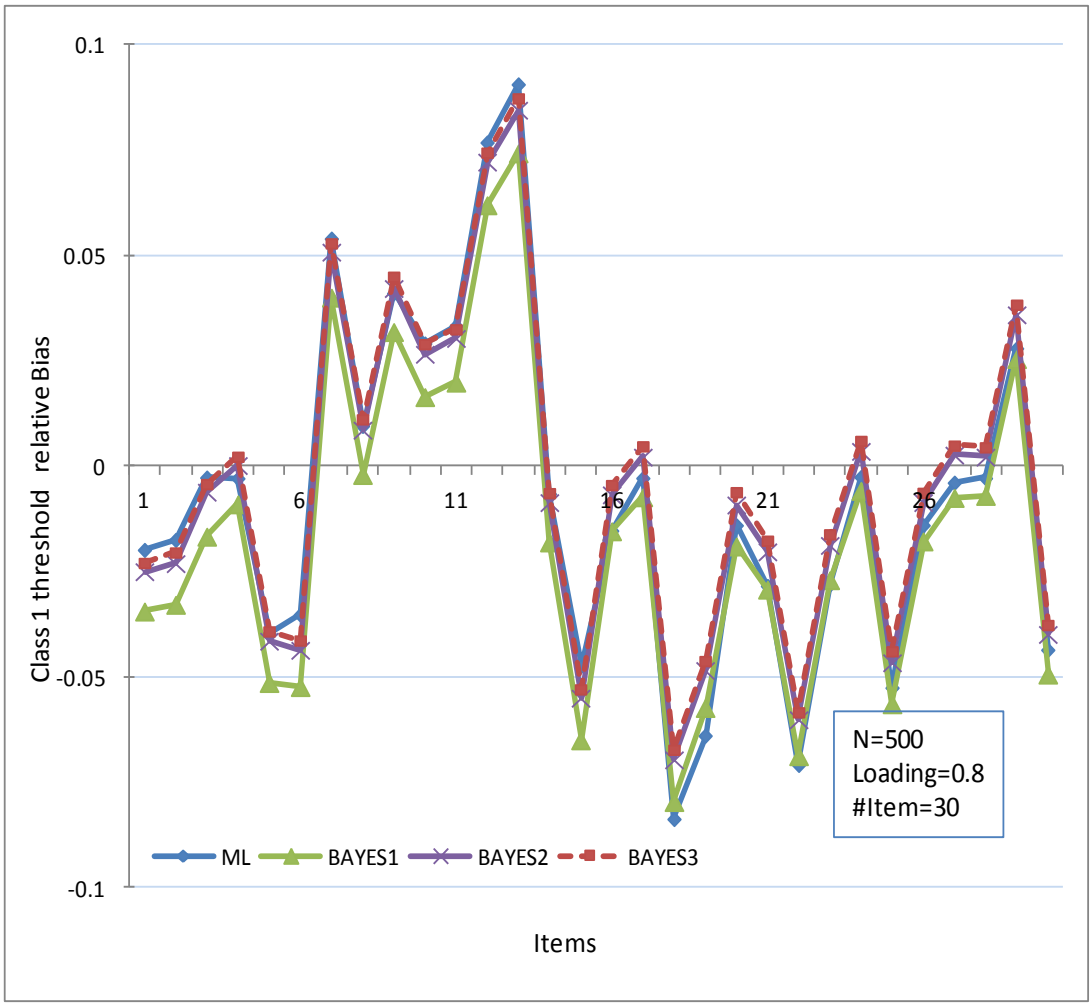


Figure 4-12. Relative bias of class 1 threshold for each item when N = 1000, loading = 0.8 and number of items = 30

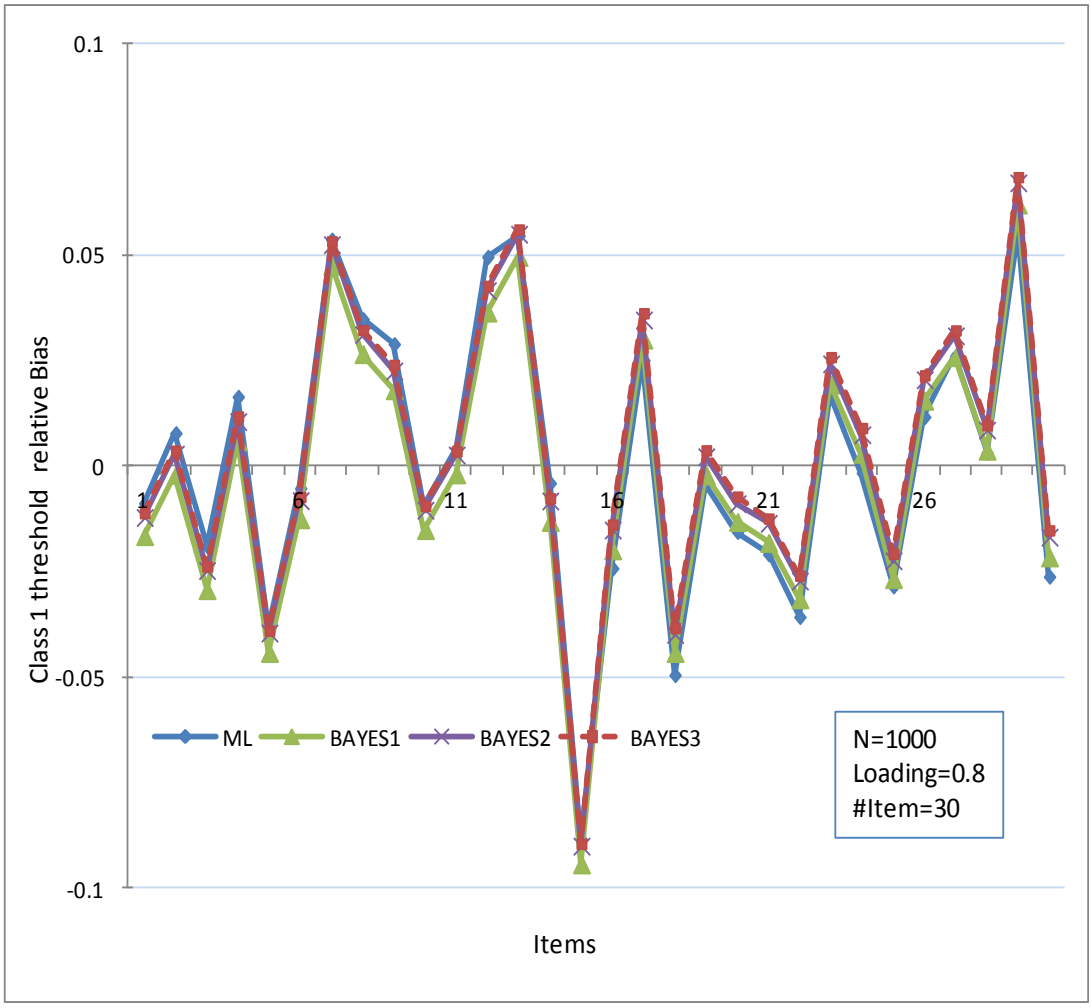


Figure 4-13. Relative bias of class 1 threshold for each item when N = 2000, loading = 0.8 and number of items = 30

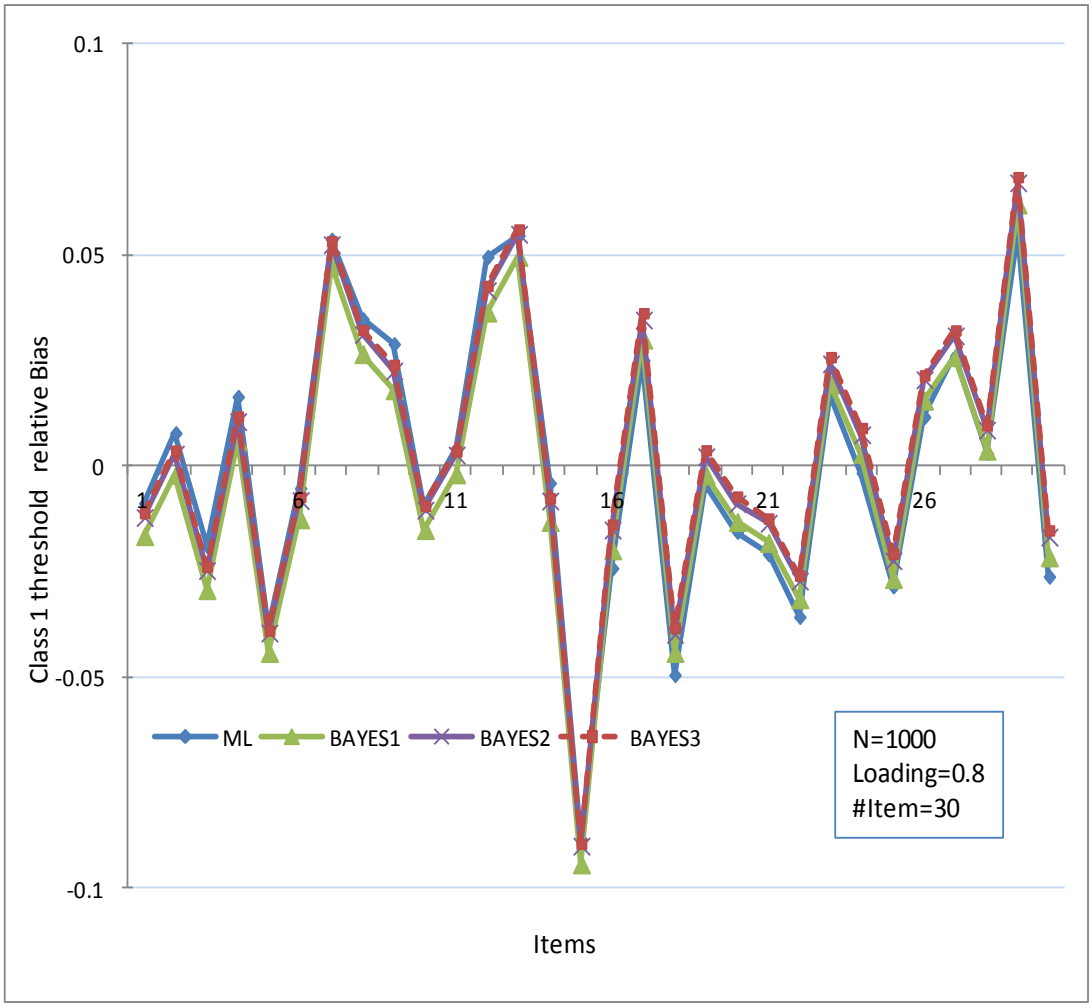
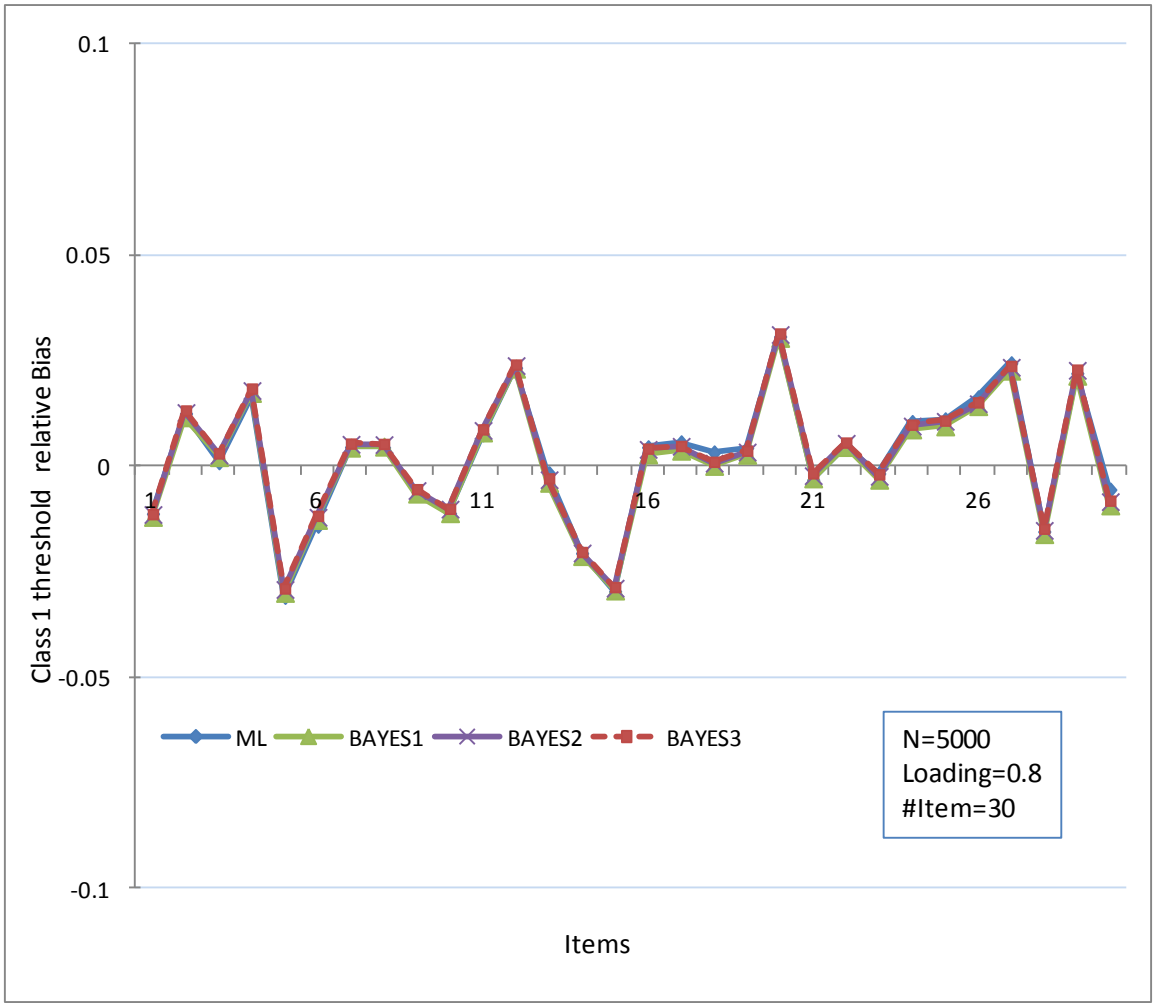


Figure 4-14. Relative bias of class 1 threshold for each item when N = 5000, loading = 0.8 and number of items = 30



Four-way ANOVA (sample size * number of items * magnitude of loading * estimation method) on the averaged Relative Bias of class 1 threshold parameters over items was conducted. Estimation method has a non-significant main effect on the relative bias of the class 1 threshold, F(3, 1492) = 1.351, p > 0.05, η_p^2= 0.003. The sample size main effect is significant, F (1, 1492) = 3.548, p < 0.05, η_p^2= 0.007. The interaction of the number of item by sample size is significant, F(3, 1492) = 2.675, p<

0.01, η_p^2= 0.012. Since the focus of the study is the comparison of estimation methods, no further analyses were conducted given the non-significant main effect and interaction related to estimation methods.

4.1.3 Recovery of Class 2 Threshold Parameters

Figures 4-15 to 4-18, B17 to B24, and B33 to B36 in Appendix B are the visual illustrations of the performance comparison among the estimation methods in terms of the relative bias of Class 2 threshold parameters. The relative biases were comparable under most conditions among all the estimation methods. Similar with the results of Class 1 threshold parameters, the difference between ML and Bayesian recovery of thresholds parameters tends to be less influenced by sample sizes compared with recovery of loading parameters, although as the sample size increases the relative biases of estimates greatly decrease and very close to each other among all the estimation methods at sample size of 2000 and 5000.

The relative biases are generally larger in the conditions with 8 items than the ones with 30 items, especially when sample size is small such as 500 and 1000 (e.g., Figure B-21 vs. B-29).

The higher loading yields lower relative bias of the class 2 threshold parameter estimates in the conditions with 8 items and a sample size of 500 (e.g. Figure B-21 vs. B-17).

Figure 4-15. Relative bias of class 2 threshold for each item when N = 500, loading = 0.8 and number of items = 30

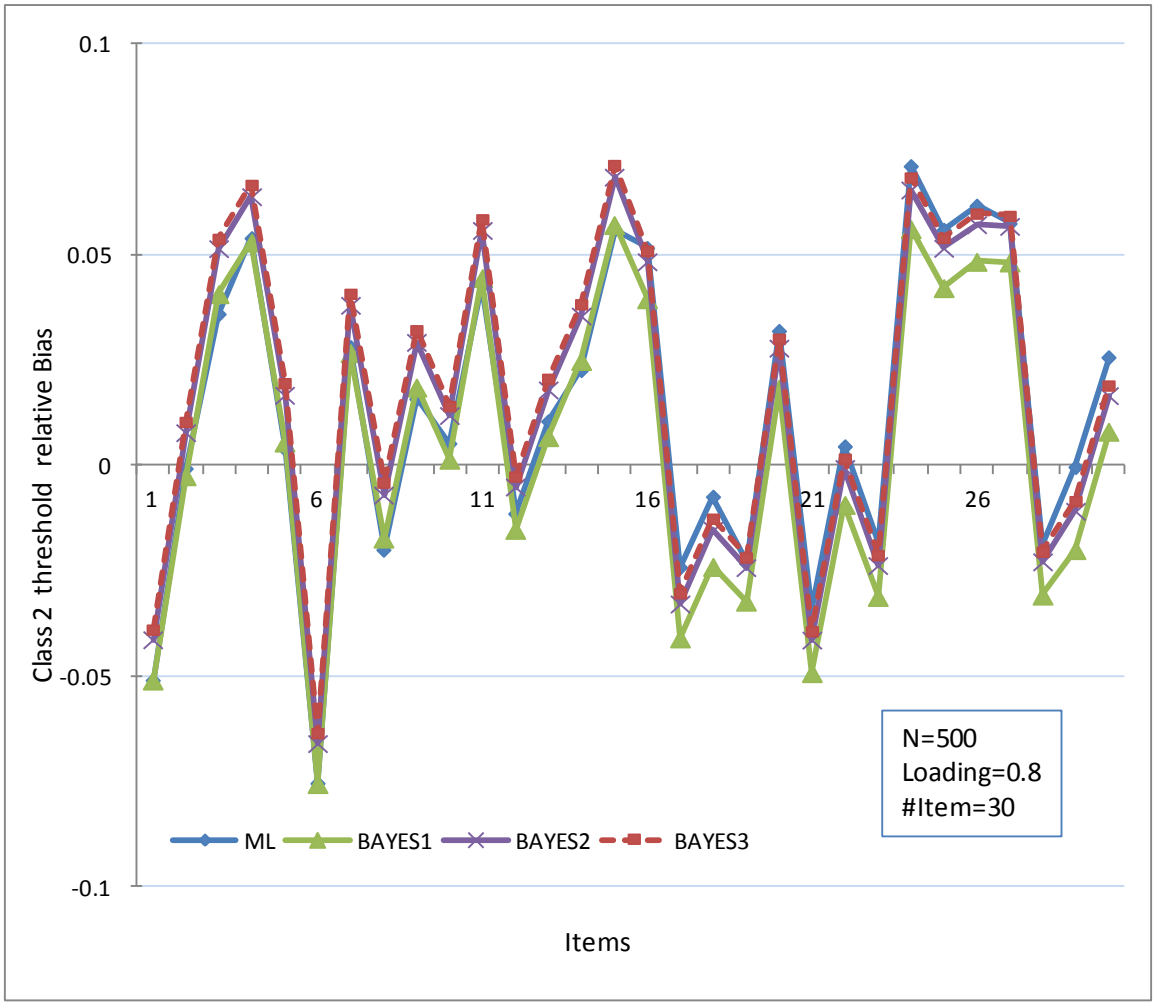


Figure 4-16. Relative bias of class 2 threshold for each item when N = 1000, loading = 0.8 and number of items = 30

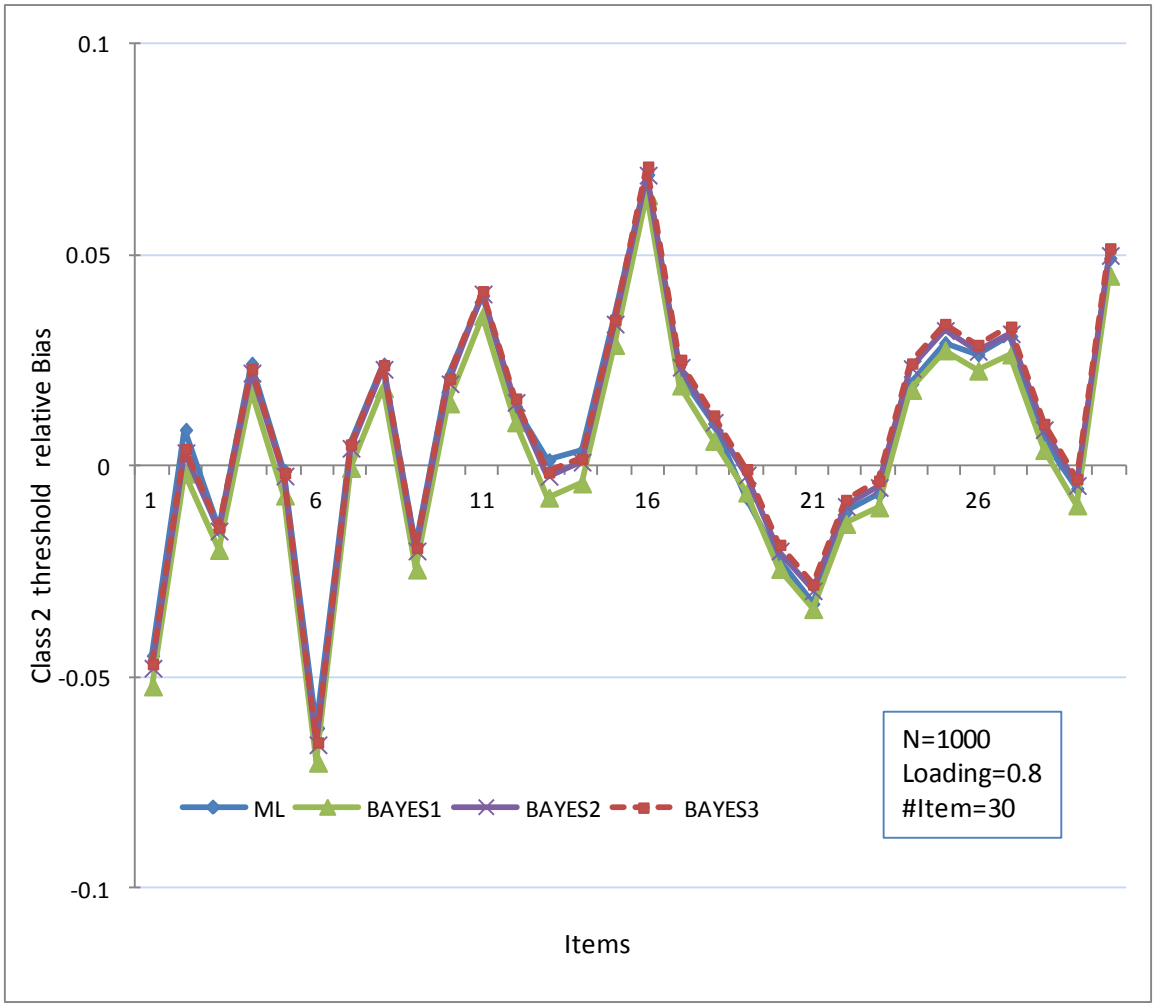


Figure 4-17. Relative bias of class 2 threshold for each item when N = 2000, loading = 0.8 and number of items = 30

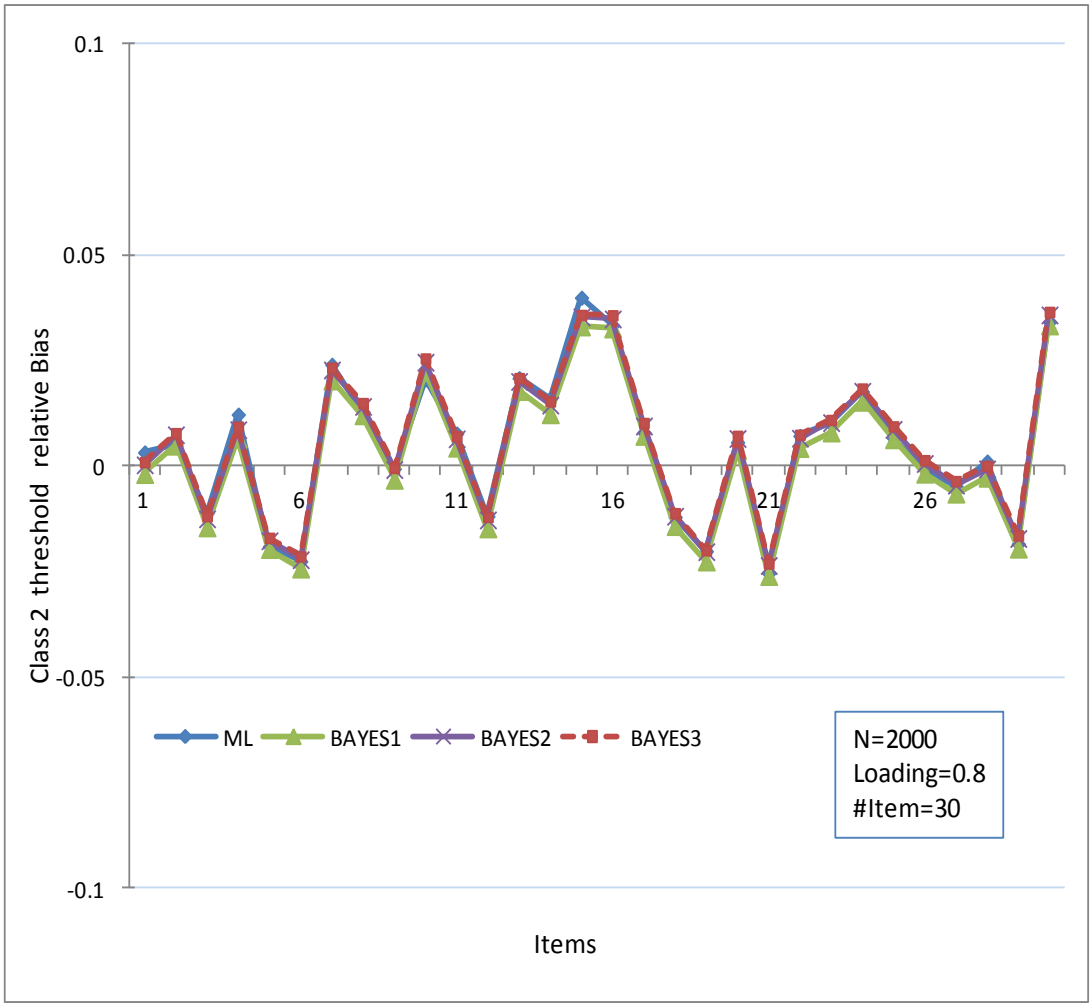
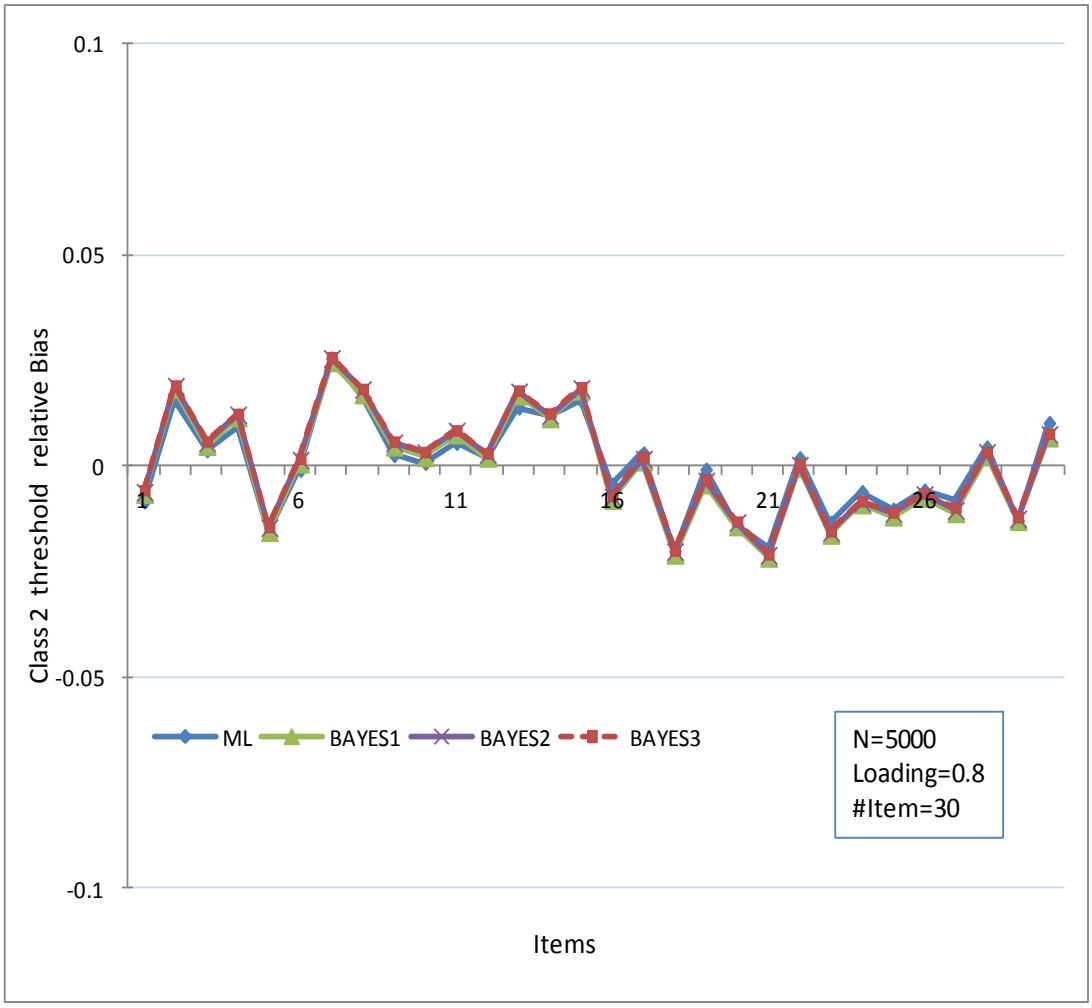


Figure 4-18. Relative bias of class 2 threshold for each item when N = 5000, loading = 0.8 and number of items = 30



Four-way ANOVA (sample size * number of items * magnitude of loading * estimation method) on the averaged Relative Bias of class 1 threshold parameters over items was conducted. Estimation method has a non-significant main effect on the relative bias of the class 2 threshold, $F(3, 1492) = 1.424$, $p > 0.05$, $\eta_p^2 = 0.003$. The main effect of the number of items is significant, $F(1, 1492) = 17.901$, $p < .01$, $\eta_p^2 = 0.012$. Again, because the focus of the study is the comparison of estimation methods,

no further analyses were conducted given the non-significant main effect and interaction associated with estimation methods.

4.2 Classification Accuracy

The mixing proportion parameters are very well estimated in all the conditions. The average of mixing proportion estimates across replications in each condition are exactly the same with the population value of 50% when they are rounded to two digits numbers.

The estimated class memberships based on class probabilities are compared to the population values in terms of average percentage of correct classification in each condition. Table 4-3 shows the average percentage of classification accuracy for each condition. The difference between all estimation methods in terms of percentage of correct classification is less than 1% for most conditions. For the conditions with 8 items, the sample size of 500 and 1000, and the magnitude of loading 0.4, the difference between all estimation methods in terms of the percentage is less than 5%. Recall that Table 4-1 represents 16 different combinations of manipulated conditions in the simulation. Figure 4-19 show that estimation methods almost result in the almost same percentage of correct classification except when there are low loading, small sample size, and small number of indicators.

Figure 4-19. Percentage of correct classification

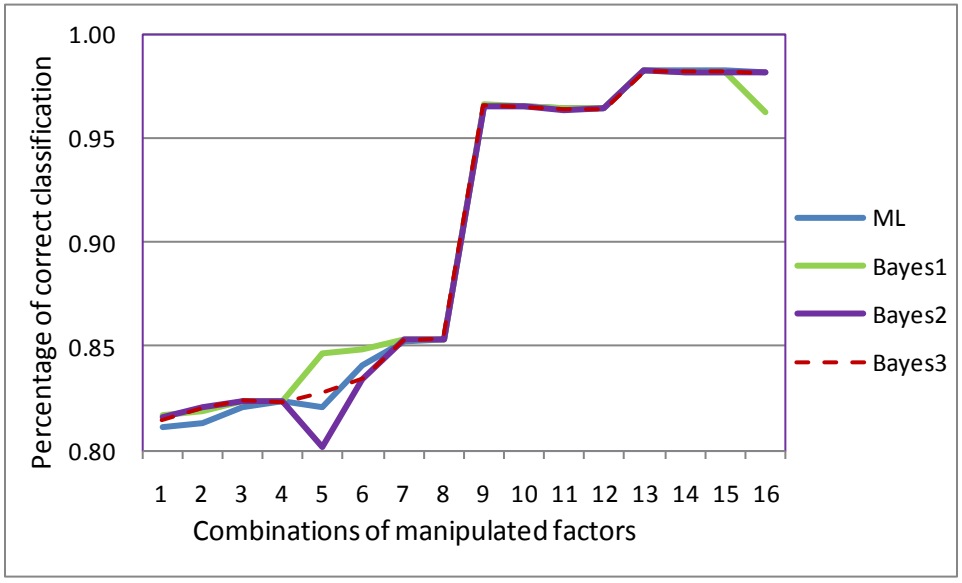


Table 4-3 Average percentage of individuals assigned to the correct latent class.

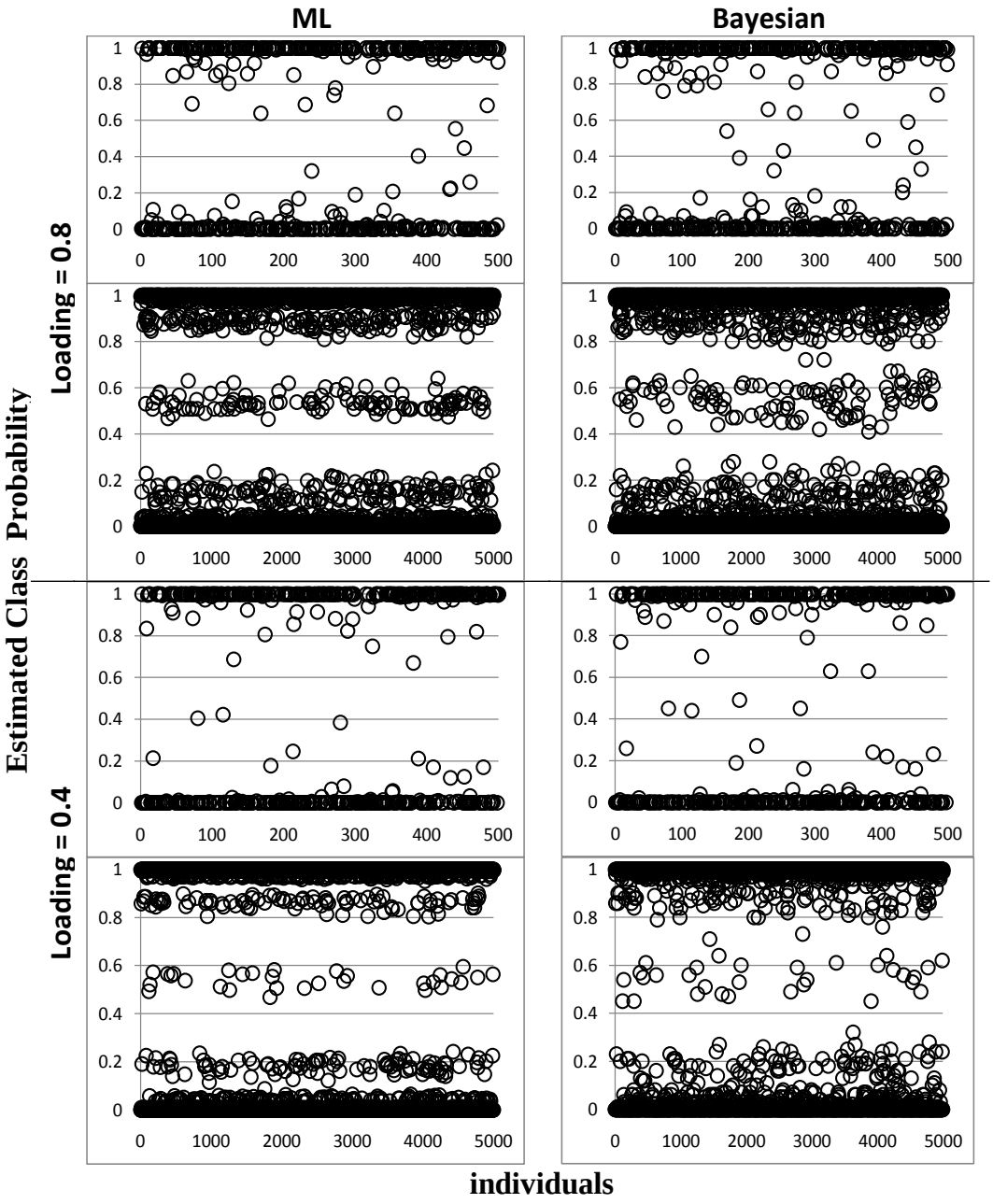
# of items	loading	sample size	ML	Bayes(0,1)	Bayes(0,5)	Bayes(0,10^5)
8	0.8	500	0.8111	0.8170	0.8157	0.8148
		1000	0.8134	0.8191	0.8204	0.8205
		2000	0.8204	0.8235	0.8235	0.8239
		5000	0.8235	0.8232	0.8235	0.8231
	0.4	500	0.8211	0.8469	0.8010	0.8278
		1000	0.8412	0.8487	0.8337	0.8342
		2000	0.8519	0.8530	0.8530	0.8529
		5000	0.8530	0.8536	0.8530	0.8536
30	0.8	500	0.9656	0.9665	0.9658	0.9661
		1000	0.9655	0.9656	0.9654	0.9653
		2000	0.9647	0.9643	0.9641	0.9640
		5000	0.9646	0.9644	0.9643	0.9642
	0.4	500	0.9829	0.9825	0.9826	0.9828
		1000	0.9825	0.9822	0.9822	0.9823
		2000	0.9825	0.9823	0.9823	0.9823
		5000	0.9822	0.9631	0.9819	0.9819

Four-way ANOVA (sample size * number of items * magnitude of loading * estimation method) on the ARCSIN square root of proportion of correct classification (Radian) was conducted and four significant effects were found: a significant main effect of the magnitude of loadings, F(1, 1492) = 610.859, p < 0 .01, η_p^2= 0.290; a significant main effect of the sample sizes, F(3, 1492) = 3.493, p < 0 .05, η_p^2= 0.007; a significant interaction of the magnitude of loadings by the number of items, F(1, 1492) = 34.883, p< 0.01, η_p^2= 0.023; and a significant interaction of the sample size by the number of items, F(3, 1492) = 13.209, p< 0.01, η_p^2= 0.026. The estimation methods do not have any significant effect on the classification accuracy, F(3, 1492) = .566, p< 0.05, η_p^2= 0.001.

Classification accuracy is also investigated by inspecting scatter-plots of the estimated class probabilities in select conditions. As representatives of all investigated conditions, eight conditions with 30 items are selected with sample sizes of 500 and 5000, loadings 0.4 and 0.8 using ML and Bayesian estimation with noninformative priors. Figure 4-20 contains the scatter plots of the class probabilities of belonging to the correct latent classes. It can be seen that ML and Bayesian perform slightly differently to some degree with different sample sizes. The points in the scatterplots falling around 0.5 in probability scale indicate difficulty in distinguishing between the latent classes, and points at extreme values of 0 and 1 indicate perfect certainty in classifying individuals. Figure 4-20 shows that when sample size is small, probability estimates are more distinguishable around probability of 0.5 in ML than those in Bayesian methods. When sample size increases to 5000, Bayesian estimation shows more spread-out pattern (i.e. more certainty)

around probability of 0.5 but is a bit less successful in moving estimates around the ideal probability values of 0 and 1 compared with ML methods.

Figure 4-20. Scatter plots of the estimated probabilities of belonging to the correct latent classes.



4.3 Convergence Rates

Convergence rates for each investigated conditions are reported in the Table 4-4. The lower convergence rates occurred in the conditions with small number of indicators and low magnitude of factor loadings. In general, larger number of indicators, higher magnitude of loadings, and larger sample size result in higher convergence rate. Given the same data set, Bayesian methods tend to have lower convergence rates than ML estimation method when there are a small number of indicators.

Table 4-4 Convergence rate

# of items	loading	sample size	ML	Bayes(0,1)	Bayes(0,5)	Bayes(0,10^5)
8	0.8	500	100	92	100	96
		1000	100	84	92	88
		2000	100	96	96	96
		5000	100	100	96	100
	0.4	500	100	64	84	72
		1000	100	92	92	96
		2000	100	96	96	96
		5000	100	100	96	100
30	0.8	500	100	100	100	100
		1000	100	100	100	100
		2000	100	100	100	100
		5000	100	100	100	100
	0.4	500	100	100	100	100
		1000	100	100	100	100
		2000	100	100	100	100
		5000	100	100	100	100

Chapter 5: Conclusion and Discussion

The current study investigates the relation between FMMs and mixture item response theory (IRT) models. The mathematical equivalence between the IRT Graded Response models and FMM with ordered categorical outcomes are proved and presented, and conversion formulas between the parameters of FMMs and mixture graded-response IRT models in probit format are provided. It is found that item discrimination parameter is equal to the value of factor loading parameters divided by residual variance, and item difficulty parameters are associated with both factor loading and threshold parameters.

In addition, Bayesian and Maximum Likelihood approaches are compared in terms of parameter recovery and classification accuracy. The comparison is based on a Monte Carlo simulation. The data were generated under FMMs and the confirmatory analyses were conducted. Besides estimation methods which include ML and Bayesian with weakly informative and noninformative priors, three other factors were manipulated in the simulation: sample size, number of outcome indicators, and magnitude of factor loadings. There are four levels of the sample size (500, 1000, 2000 and 5000), two levels of the magnitude of factor loading (0.4 and 0.8), and two levels of the number of items (8 and 30).

5.1 Summary of the Simulation Results

ML and Bayesian estimation methods perform differently in some investigated conditions with respect to recovering item parameters and classifying respondents in the manipulated conditions of different combinations of sample size,

the number of items and magnitude of factor loadings. In general, for the investigated conditions, ML and Bayesian with weakly informative priors perform well with small sample size, and that all estimation methods perform well with large sample size. The detailed findings on comparison of the estimation methods are as follows:

(1) Estimation methods perform significantly differently with the different sample sizes for loading parameter estimates, though the effect size index indicates it is a small effect. When the sample size is small, the ML tends to have less-biased estimates than the Bayesian method using noninformative prior. As the sample size increases, the superiority of the ML over the Bayesian method decreases. The estimation methods have no significant impact on the estimates of threshold parameters.

(2) When the sample size is small, though the Bayesian estimations with two weakly informative priors do not perform significantly differently from each other, more informative prior is helpful to some degree in reducing the relative bias of the loading estimate. However, the threshold parameters and classification accuracy is not associated with how informative the prior is in Bayesian estimation as much as loading parameter does.

(3) When sample size is small, though Bayesian methods with weakly informative priors don't yield significantly lower relative bias than Bayesian with noninformative prior, weakly informative priors help reduce the relative bias to some degree for loading parameters. Threshold parameter recovery and classification accuracy are not associated with

whether Bayesian prior is informative or noninformative as much as loading parameter does.

(4) As sample size increases, the effect of the priors on Bayesian estimation tends to be negligible, rendering it comparable to ML in terms of the relative bias of parameter estimates. The relative bias of the estimates from all estimation methods are very close to each other when sample size is large.

(5) Magnitude of loadings does not have significant impact on relative bias of estimates and percentage of correct classification. Estimation methods and magnitude of loadings do not interact each other. The change of magnitude of loading does not make estimation methods perform differently from each other.

(6) Though the number of indicators has a significant effect on the relative bias and percentage of correct classification, there is no significant interaction effect between the number of binary indicators and estimation methods. In other words, the estimation method has no impacts on the relative bias of the estimates and percentage of correct classification when the number of indicators changes.

(7) The estimates are influenced by the estimation method with respect to the stability of the estimates when there are a small number of indicators. Specifically, with high magnitude of loadings the estimates from ML methods are less stable with a small number of indicators than Bayesian methods; with low magnitude of loadings, the estimates from most

informative priors are most stable, followed by ML estimates, and Bayesian estimates with noninformative priors have the largest *SE*s. As the number of indicators increasing, there are no differences among the estimation methods with respect to the stability of the estimates.

(8) Classification accuracy is not associated with the estimation method under the investigated conditions.

It is also found that there are some factors which do not have interaction effects with estimation methods but influence the relative bias and classification accuracy. First, the sample size and the number of items have effects on threshold parameter estimation. Second, the magnitude of loadings which interacts with the number of items has large effect on percentage of correct classification. Third, the sample size interacting with the number of items has significant effect on percentage of correct classification.

Flipping of the loading sign in ML often happens when the magnitude of loading is low. The flipping never happened in Bayesian estimation in the investigated conditions considered in this study. In a simulation study, ignoring it may result in exaggeration of ML biases of loading parameters when aggregating estimates across replications. Comparing aggregated results of ML with Bayesian method without controlling sign flipping may lead to invalid conclusions.

5.2 Limitation, Recommendations and Future Study

The current simulation study was limited to the mixture latent classes with equal mean structure, which is sometimes the case when studying the DIF. More often, latent groups display different latent trait distributions, thus conditions in which latent classes have different mean structure would be worth investigating.

Some population values of the model parameters are chosen at extreme cases in the current study. For example, mixing proportion is 50% to 50%. The threshold parameters are very neatly varying across classes. The manner latent classes vary is sometimes more complicated with real data than the simulated conditions in the current study. Different mixing proportion and pattern of invariant or non-invariant thresholds and loadings (or item discrimination and difficulty parameters) across classes are also possible investigated factors in the future study. In addition, the starting values in the simulation are set to the population values, which may lead to better solutions than what researchers typically find in practice.

In the current study, Bayesian estimations yield different convergence rates from ML method, and Bayesian methods with different priors also lead to different convergence rates. The convergence has often been a challenge for mixture modeling. It is interesting to investigate further how the convergence is influenced by possible factors. Increasing the number of iterations may improve the convergence rate of Bayesian methods in some conditions. Further, well-chosen priors may help convergence in Bayesian estimation. A more systematically-designed simulation may be worth conducting on how factors such as priors impact on the convergence rate.

Sign flipping of loading parameters in FMMs are controlled in the current study by constraining the values to be positive. By investigating one out of 64 manipulated conditions, it was found that this constraint on loading parameters did not have any impact on the estimates of threshold parameters. It would be meaningful to study whether controlling loading sign flipping has any impact on the estimates of latent abilities and other item parameters in various simulation conditions.

It is found that besides estimation methods some factors including interaction of the factors may have impact on parameter estimates and individual classification. Those are important to study in detail for the practitioners in the future study.

Finally, based on the result from the current simulation it is recommended for FMM or mixture IRT modelers that

(1) Simulation study should always be conducted before applying specific factor mixture or mixture IRT models.

(2) If sample size is small, ML or Bayesian with weakly informative priors methods is strongly recommended.

(3) Weakly informative priors generally perform better than noninformative priors, thus are recommended in the FMMs or Mixture IRT modeling with binary outcomes even though when we don't have prior knowledge of the values of model parameters, given the special characteristics of the probit and logit regression.

(4) When sample size is large enough such as 5000, both ML and Bayesian methods perform well in parameter recovery and classification.

(5) In simulations, the flipping of the loading signs need to be controlled in ML estimation. Otherwise, ML bias for loading parameters would be overinflated. It is less likely to happen in Bayesian estimation.

(6) With low loading/discrimination and small sample size, FMM/Mixture IRT not recommended.

(7) For both ML and Bayesian methods, larger number of indicators, higher magnitude of loadings, and larger sample size result in higher convergence rates. To deal with non-convergence for a given set of data, if a change of convergence criterion is not feasible, one may increase the number of iterations, or try different starting values. If both do not work, the model may need to be modified such as adding more constraints on latent class membership.

Appendix A

Table A-1 Bias, Relative Bias and Standard Error of Loading Estimates

Loading			ML			Bayes (0,1)			Bayes(0,5)			Bayes(0, 10^5)		
# of items	loading	sample size	bias	relative bias	*SE*	Bias	relative bias	*SE*	bias	relative bias	*SE*	bias	relative bias	*SE*
8 items	0.8	500	0.0040	0.0050	0.0666	0.0054	0.0067	0.0610	0.0197	0.0246	0.0626	0.0276	0.0345	0.0631
		1000	-0.0127	-0.0158	0.0612	-0.0034	-0.0042	0.0491	0.0049	0.0062	0.0483	0.0071	0.0088	0.0465
		2000	-0.0041	-0.0051	0.0366	-0.0016	-0.0020	0.0288	0.0011	0.0013	0.0294	0.0020	0.0025	0.0282
		5000	0.0000	0.0000	0.0202	-0.0011	-0.0013	0.0194	0.0011	0.0013	0.0294	0.0004	0.0006	0.0194
	0.4	500	-0.0164	-0.0410	0.0838	0.0082	0.0204	0.0644	-0.0016	-0.0040	0.1104	0.0690	0.1724	0.2451
		1000	-0.0248	-0.0423	0.0704	-0.0037	-0.0246	0.0486	-0.0005	-0.0013	0.0492	0.0000	-0.0001	0.0494
		2000	-0.0045	-0.0112	0.0333	-0.0060	-0.0151	0.0294	-0.0033	-0.0083	0.0293	-0.0031	-0.0076	0.0292
		5000	-0.0001	-0.0004	0.0214	-0.0011	-0.0029	0.0212	-0.0033	-0.0083	0.0293	-0.0005	-0.0012	0.0212
30 items	0.8	500	0.0025	0.0031	0.0171	0.0130	0.0163	0.0172	0.0341	0.0426	0.0179	0.0398	0.0498	0.0181
		1000	0.0000	0.0000	0.0117	0.0059	0.0074	0.0117	0.0158	0.0198	0.0119	0.0184	0.0230	0.0120
		2000	-0.0041	-0.0051	0.0084	-0.0029	-0.0036	0.0084	0.0015	0.0019	0.0085	0.0029	0.0037	0.0085
		5000	0.0000	0.0000	0.0052	0.0011	0.0014	0.0052	0.0027	0.0034	0.0052	0.0033	0.0041	0.0053
	0.4	500	-0.0002	-0.0004	0.0153	0.0107	0.0267	0.0157	0.0151	0.0378	0.0159	0.0163	0.0407	0.0160
		1000	-0.0020	-0.0049	0.0105	0.0035	0.0087	0.0107	0.0055	0.0138	0.0107	0.0060	0.0150	0.0107
		2000	-0.0038	-0.0095	0.0076	-0.0017	-0.0043	0.0076	-0.0008	-0.0020	0.0077	-0.0005	-0.0014	0.0077
		5000	-0.0005	-0.0014	0.0048	0.0004	0.0010	0.0048	0.0008	0.0020	0.0048	0.0009	0.0023	0.0048

Table A-2 *Bias, Relative Bias and Standard Error of Class 1 Threshold Estimates*

Class 1 threshold			ML			Bayes (0,1)			Bayes(0,5)			Bayes(0, 10^5)		
# of items	loading	sample size	bias	relative bias	*SE*	Bias	relative bias	*SE*	bias	relative bias	*SE*	bias	relative bias	*SE*
8 items	0.8	500	0.0071	0.0197	0.0776	0.0138	0.0046	0.0651	0.0028	0.0136	0.0665	0.0129	0.0247	0.0682
		1000	-0.0298	-0.0391	0.0730	-0.0142	-0.0296	0.0536	-0.0190	-0.0217	0.0549	-0.0148	-0.0323	0.0496
		2000	-0.0082	-0.0120	0.0457	0.0215	-0.0155	0.0353	0.0163	-0.0095	0.0359	0.0206	-0.0075	0.0343
		5000	0.0066	0.0010	0.0260	0.0009	-0.0017	0.0244	0.0163	-0.0095	0.0359	0.0010	0.0002	0.0245
	0.4	500	-0.0097	-0.0419	0.0826	-0.0211	0.0102	0.0508	0.0078	-0.0345	0.0787	-0.0426	0.0989	0.1391
		1000	0.0140	-0.0367	0.0518	0.0025	-0.0532	0.0450	0.0013	-0.0314	0.0453	-0.0016	-0.0287	0.0450
		2000	-0.0006	-0.0038	0.0304	0.0050	-0.0176	0.0272	0.0079	-0.0091	0.0272	0.0079	-0.0083	0.0272
		5000	-0.0019	-0.0115	0.0168	-0.0043	-0.0134	0.0167	0.0079	-0.0091	0.0272	-0.0043	-0.0122	0.0167
30 items	0.8	500	0.0094	-0.0078	0.0044	0.0064	-0.0153	0.0044	0.0066	-0.0056	0.0045	0.0065	-0.0032	0.0045
		1000	0.0024	0.0004	0.0036	-0.0005	-0.0020	0.0036	-0.0004	0.0027	0.0036	-0.0006	0.0038	0.0037
		2000	0.0030	0.0037	0.0031	0.0044	0.0012	0.0031	0.0045	0.0034	0.0031	0.0044	0.0041	0.0031
		5000	-0.0028	0.0019	0.0024	-0.0024	0.0012	0.0024	-0.0025	0.0020	0.0024	-0.0025	0.0023	0.0024
	0.4	500	0.0134	-0.0086	0.0041	0.0119	-0.0136	0.0042	0.0121	-0.0070	0.0042	0.0121	-0.0053	0.0042
		1000	0.0044	-0.0031	0.0034	0.0036	-0.0047	0.0034	0.0037	-0.0013	0.0035	0.0037	-0.0006	0.0035
		2000	0.0010	0.0032	0.0029	0.0012	0.0017	0.0029	0.0013	0.0033	0.0029	0.0013	0.0038	0.0029
		5000	-0.0014	0.0026	0.0023	-0.0014	0.0021	0.0023	-0.0014	0.0028	0.0023	-0.0014	0.0029	0.0023

Table A-3 *Bias, Relative Bias and Standard Error of Class2 Threshold Estimates*

Class 2 threshold			ML			Bayes (0,1)			Bayes(0,5)			Bayes(0, 10^5)		
# of items	loading	sample size	bias	relative bias	*SE*	Bias	relative bias	*SE*	bias	relative bias	*SE*	bias	relative bias	*SE*
8 items	0.8	500	0.0346	-0.0038	0.0802	0.0189	-0.0328	0.0656	0.0383	-0.0073	0.0692	0.0279	0.0041	0.0667
		1000	0.0520	-0.0195	0.0671	0.0334	-0.0132	0.0515	0.0387	-0.0077	0.0496	0.0377	0.0006	0.0501
		2000	0.0170	-0.0092	0.0453	-0.0148	-0.0134	0.0321	-0.0080	-0.0102	0.0330	-0.0122	-0.0098	0.0311
		5000	-0.0044	-0.0024	0.0248	0.0015	-0.0078	0.0231	-0.0080	-0.0102	0.0330	0.0014	-0.0061	0.0231
	0.4	500	0.0390	-0.0604	0.0710	0.0450	-0.0215	0.0498	0.0070	-0.0571	0.0822	0.0543	0.0244	0.1280
		1000	-0.0105	-0.0006	0.0518	0.0090	-0.0097	0.0373	0.0101	-0.0010	0.0377	0.0152	-0.0006	0.0385
		2000	0.0084	-0.0160	0.0263	0.0031	-0.0310	0.0243	0.0033	-0.0171	0.0242	0.0032	-0.0166	0.0242
		5000	0.0057	-0.0001	0.0169	0.0081	-0.0049	0.0168	0.0033	-0.0171	0.0242	0.0081	-0.0035	0.0168
30 items	0.8	500	0.0019	0.0117	0.0022	-0.0016	0.0047	0.0022	-0.0021	0.0143	0.0022	-0.0021	0.0169	0.0022
		1000	0.0023	0.0072	0.0020	0.0033	0.0026	0.0020	0.0033	0.0072	0.0020	0.0034	0.0084	0.0020
		2000	-0.0009	0.0047	0.0018	-0.0006	0.0025	0.0018	-0.0006	0.0048	0.0018	-0.0006	0.0055	0.0018
		5000	-0.0032	0.0002	0.0016	-0.0041	-0.0003	0.0016	-0.0041	0.0004	0.0016	-0.0041	0.0007	0.0016
	0.4	500	-0.0004	0.0109	0.0021	-0.0013	0.0064	0.0022	-0.0013	0.0134	0.0022	-0.0013	0.0152	0.0022
		1000	0.0001	0.0021	0.0019	0.0001	-0.0011	0.0020	0.0001	0.0022	0.0020	0.0001	0.0030	0.0020
		2000	-0.0015	0.0014	0.0018	-0.0013	-0.0002	0.0018	-0.0013	0.0015	0.0018	-0.0013	0.0019	0.0018
		5000	-0.0020	-0.0016	0.0016	-0.0022	-0.0021	0.0016	-0.0022	-0.0014	0.0016	-0.0022	-0.0012	0.0016

Appendix B

Figure B-1. Relative bias of loading for each item when N = 500, loading = 0.8 and number of items = 8

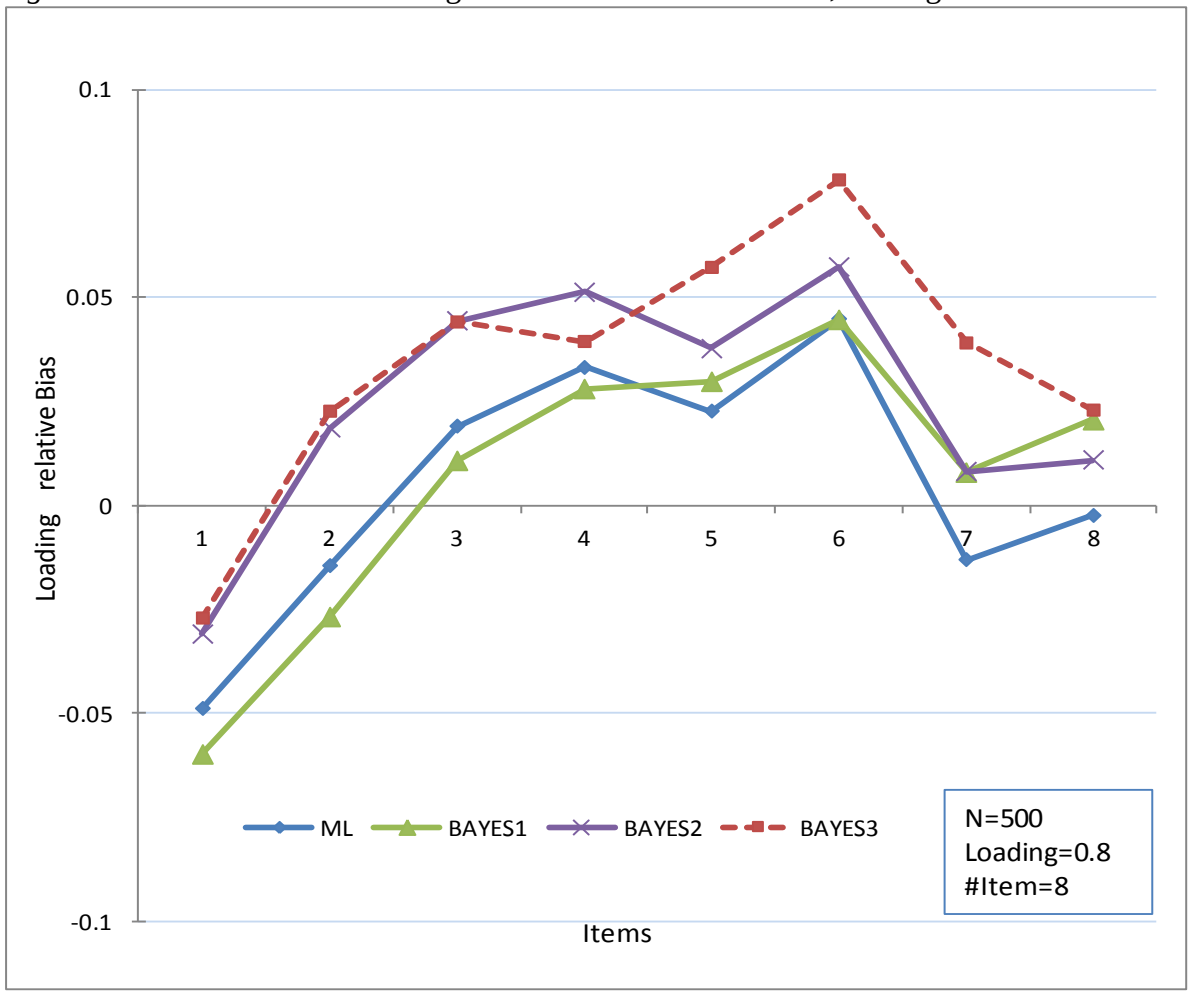


Figure B-2. Relative bias of loading for each item when N = 1000, loading = 0.8 and number of items = 8

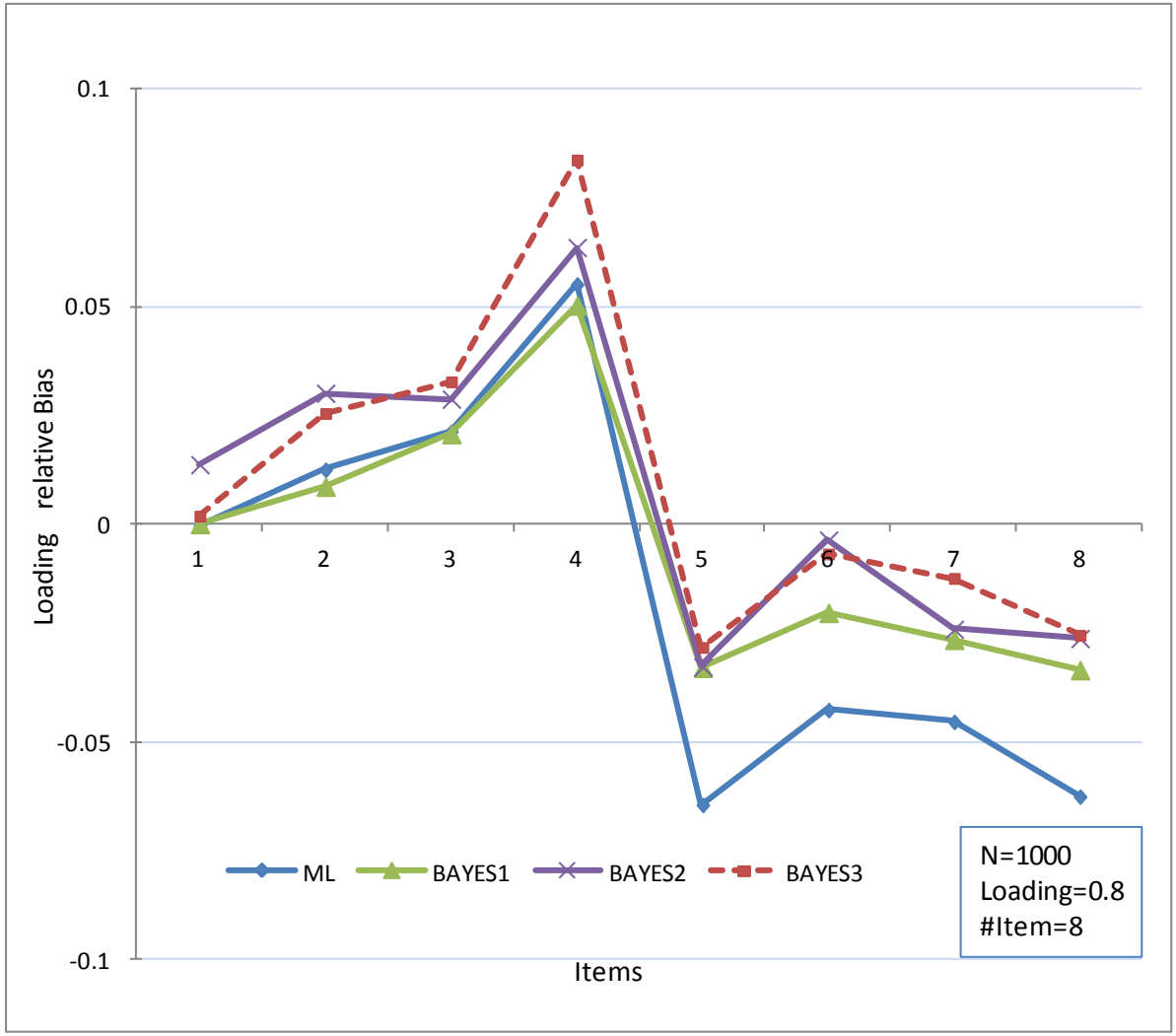


Figure B-3. Relative bias of loading for each item when N = 2000, loading= 0.8 and number of items = 8

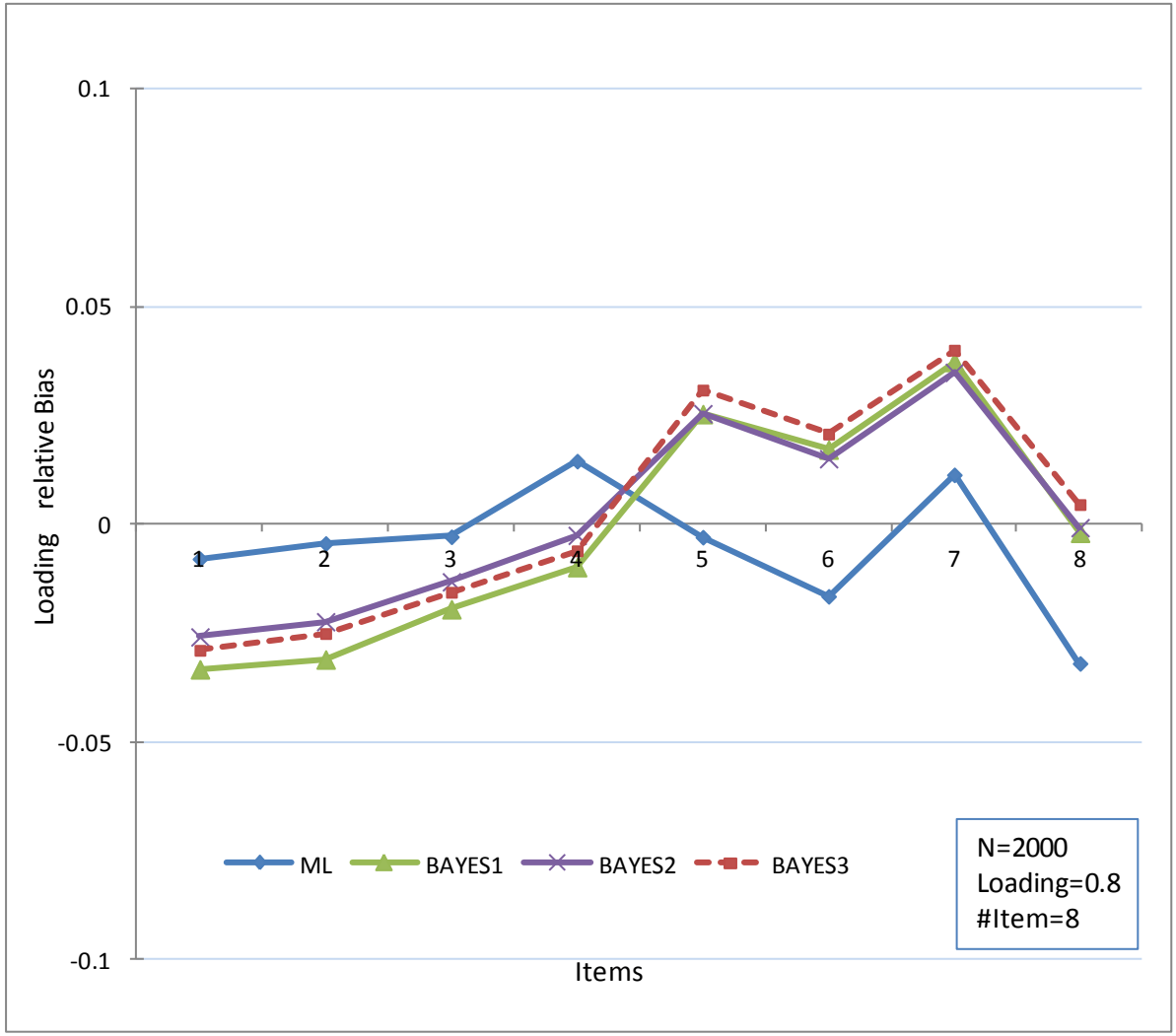


Figure B-4. Relative bias of loading for each item when N = 5000, loading = 0.8 and number of items = 8

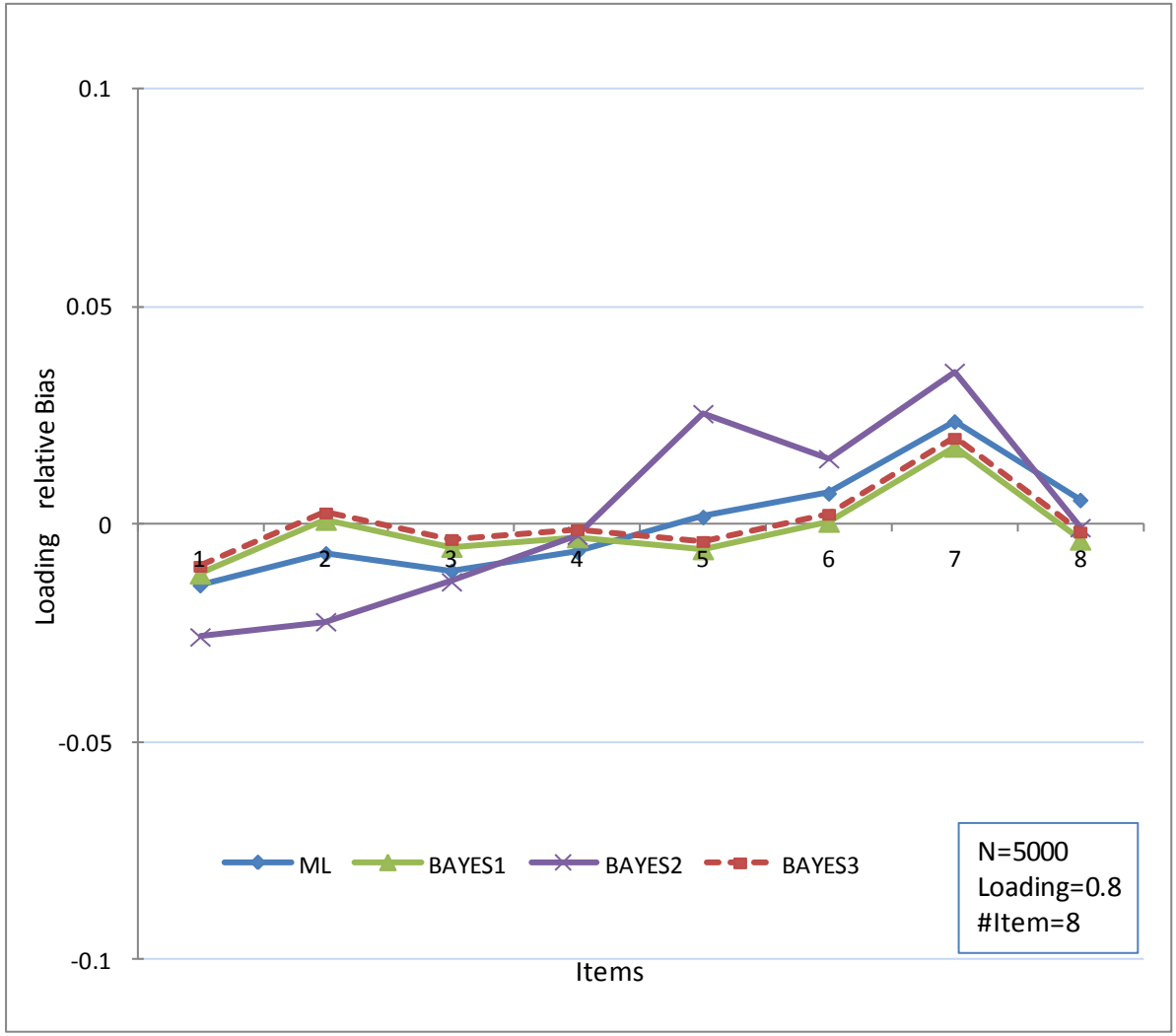


Figure B-5. Relative bias of loading for each item when N = 500, loading = 0.4 and number of items = 8

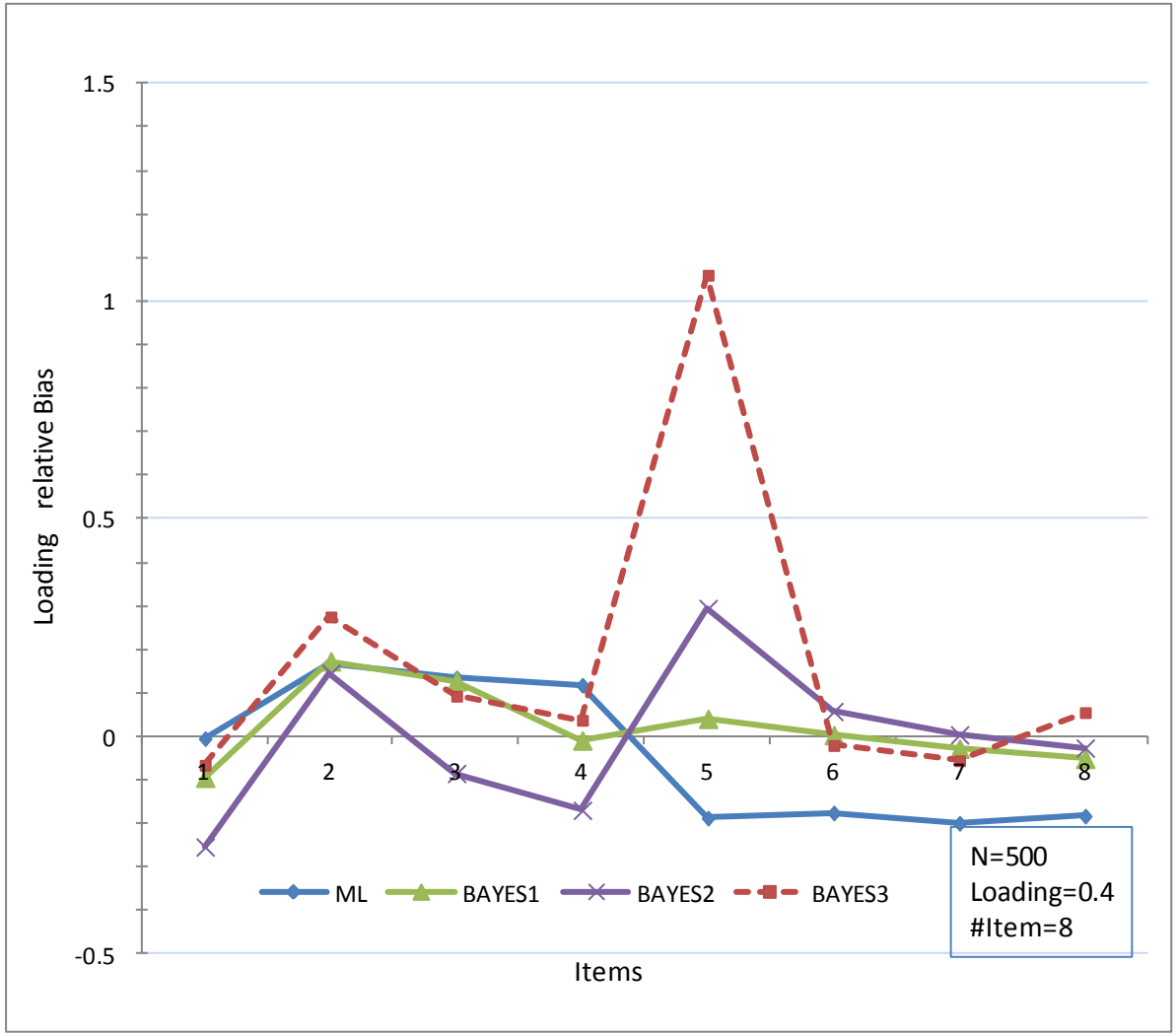


Figure B-6. Relative bias of loading for each item when N = 1000, loading = 0.4 and number of items = 8

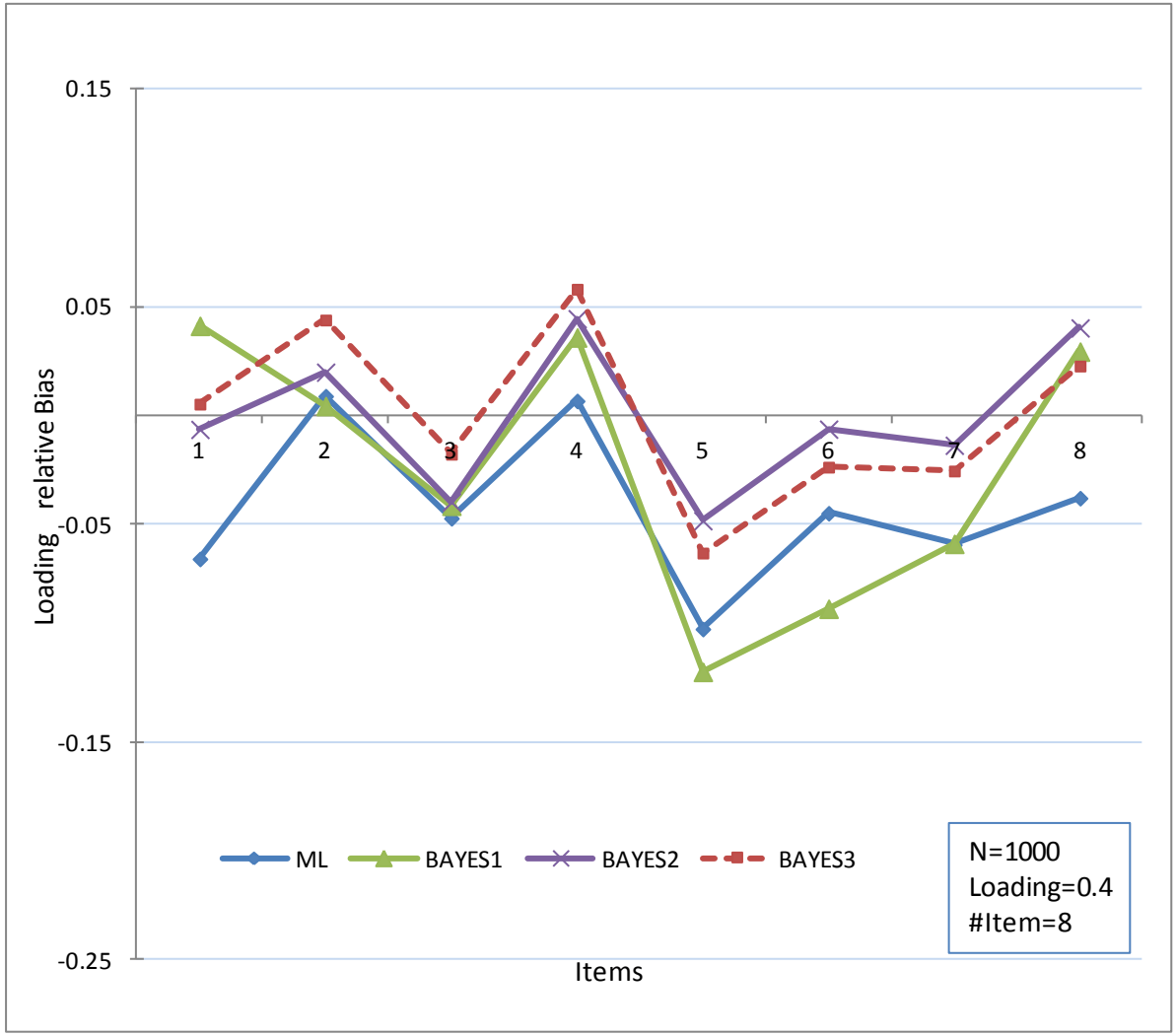


Figure B-7. Relative bias of loading for each item when N = 2000, loading = 0.4 and number of items = 8

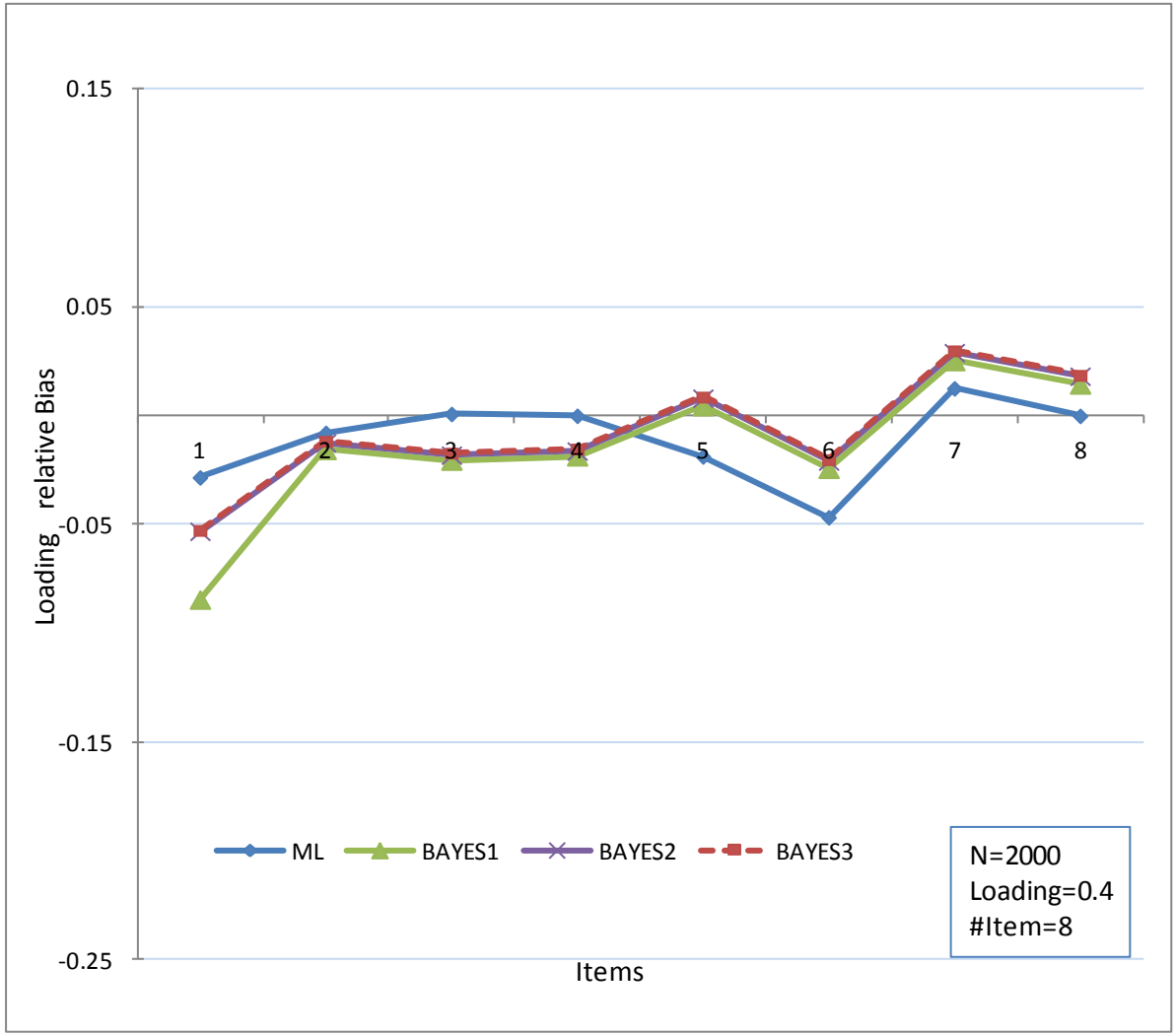


Figure B-8. Relative bias of loading for each item when N = 5000, loading = 0.4 and number of items = 8

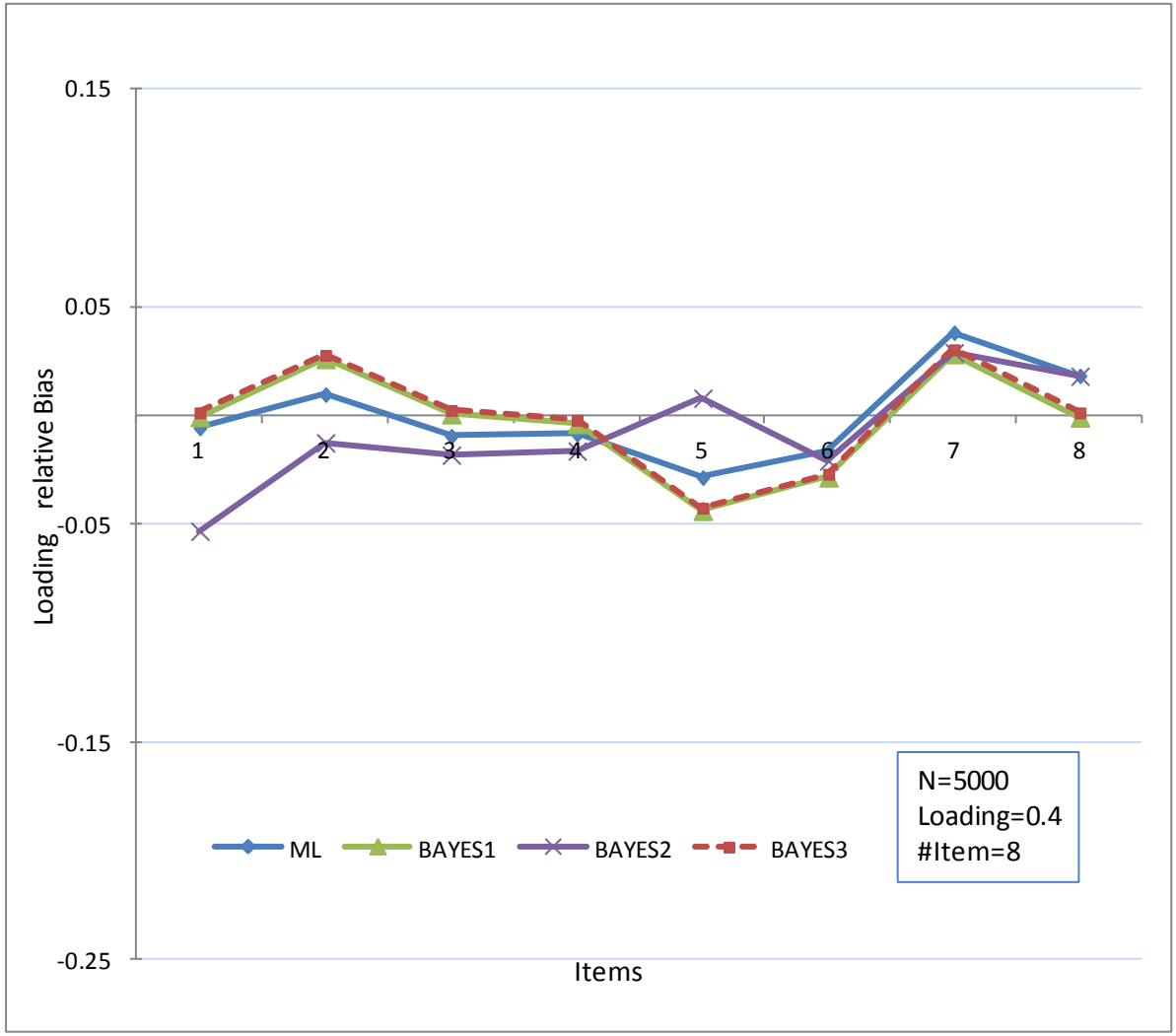


Figure B-9. Relative bias of class1 threshold for each item when N = 500, loading = 0.8 and number of items = 8

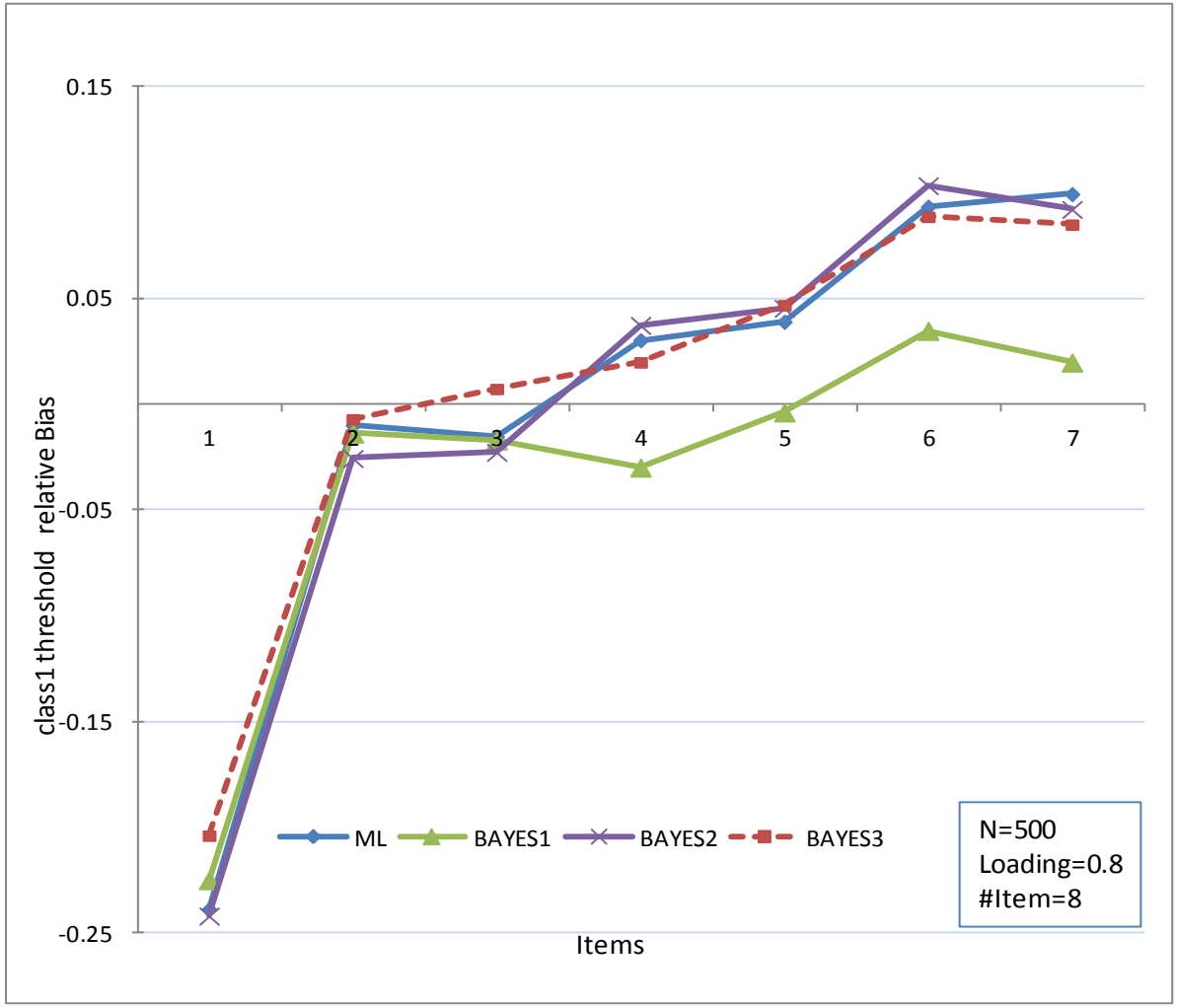


Figure B-10. Relative bias of class 1 threshold for each item when N = 1000, loading = 0.8 and number of items = 8

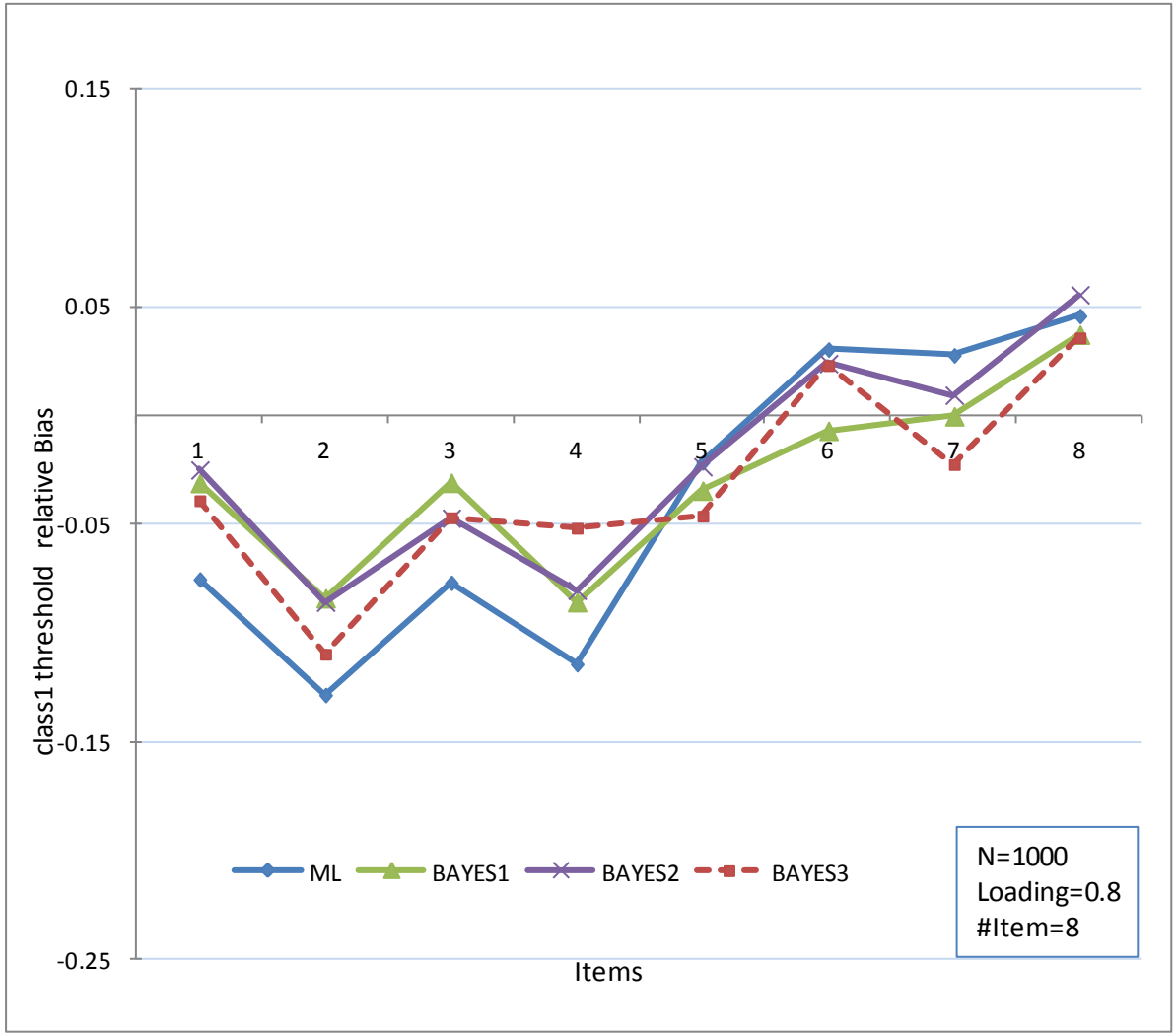


Figure B-11. Relative bias of class 1 threshold for each item when N = 2000, loading = 0.8 and number of items = 8

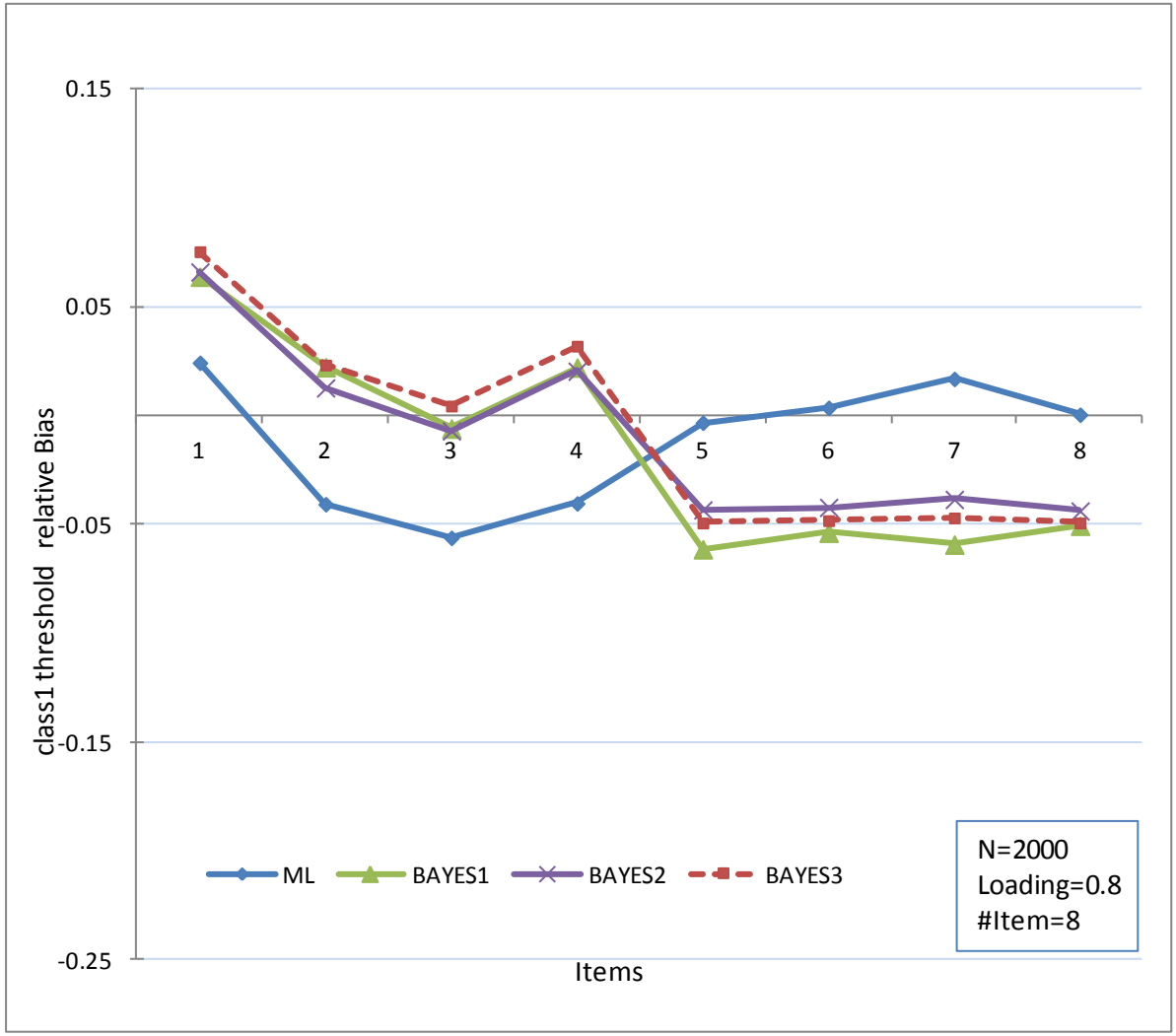


Figure B-12. Relative bias of class 1 threshold for each item when N = 5000, loading = 0.8 and number of items = 8

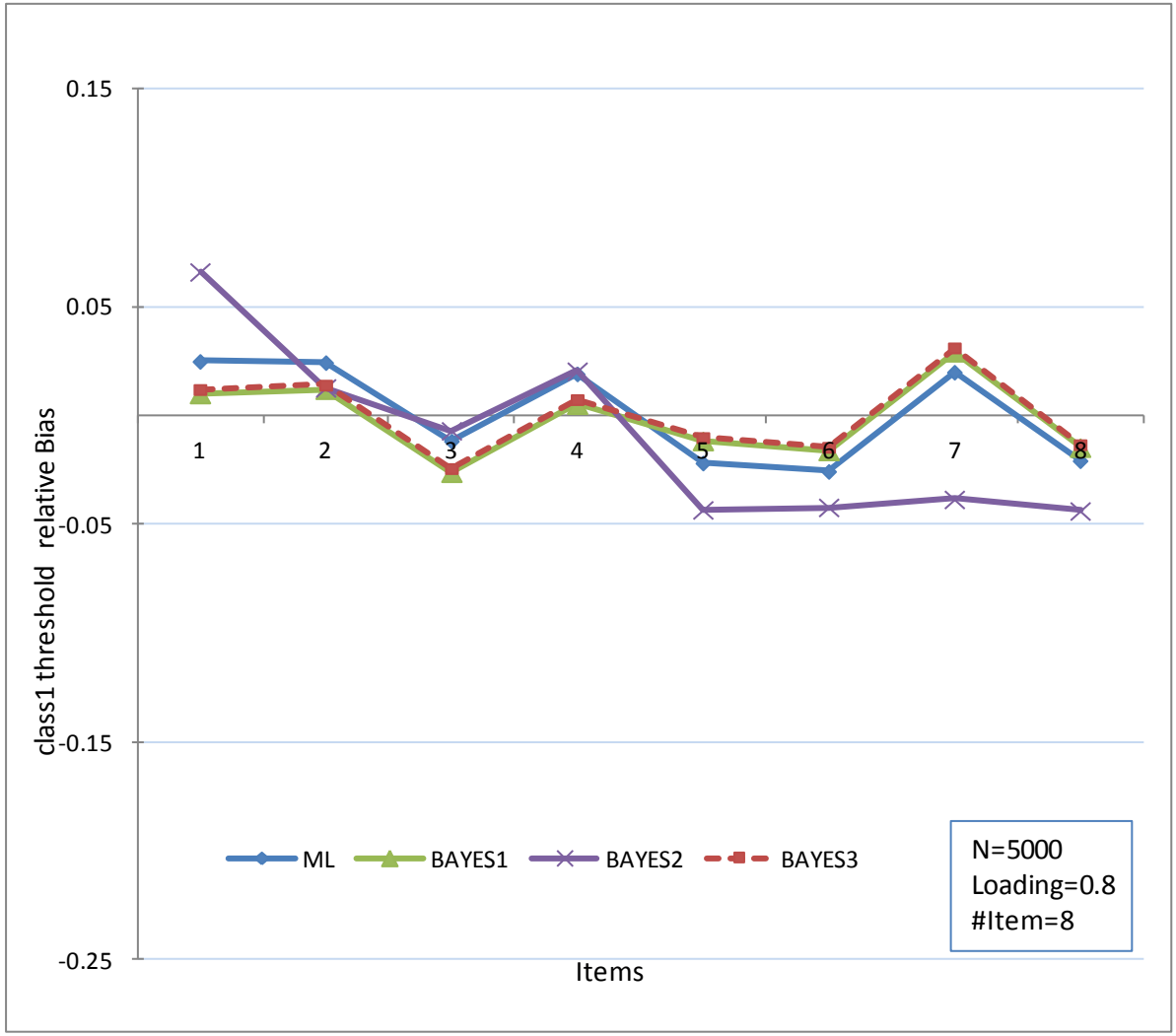


Figure B-13. Relative bias of class 1 threshold for each item when N = 500, loading = 0.4 and number of items = 8

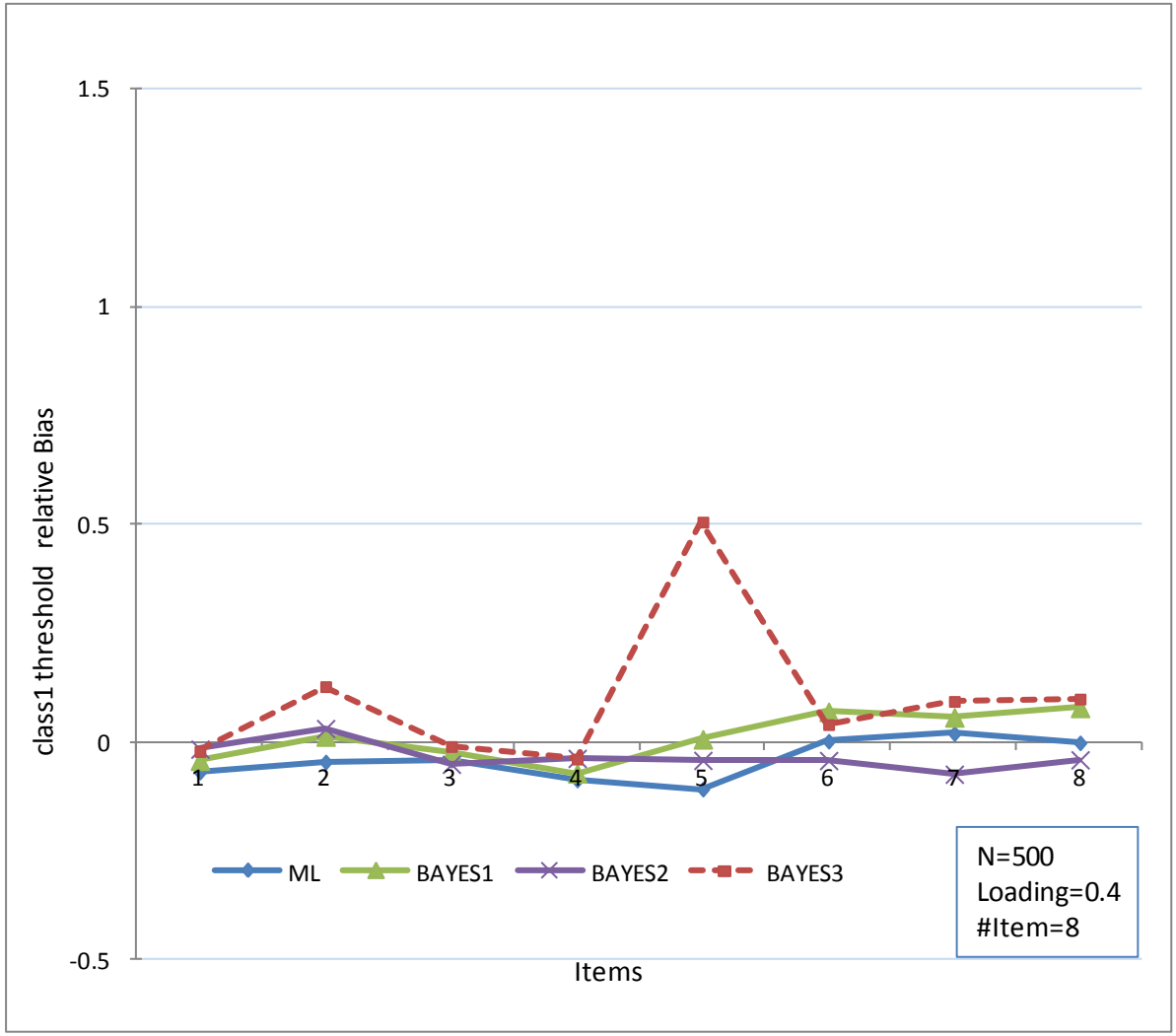


Figure B-14. Relative bias of class 1 threshold for each item when N = 1000, loading = 0.4 and number of items = 8

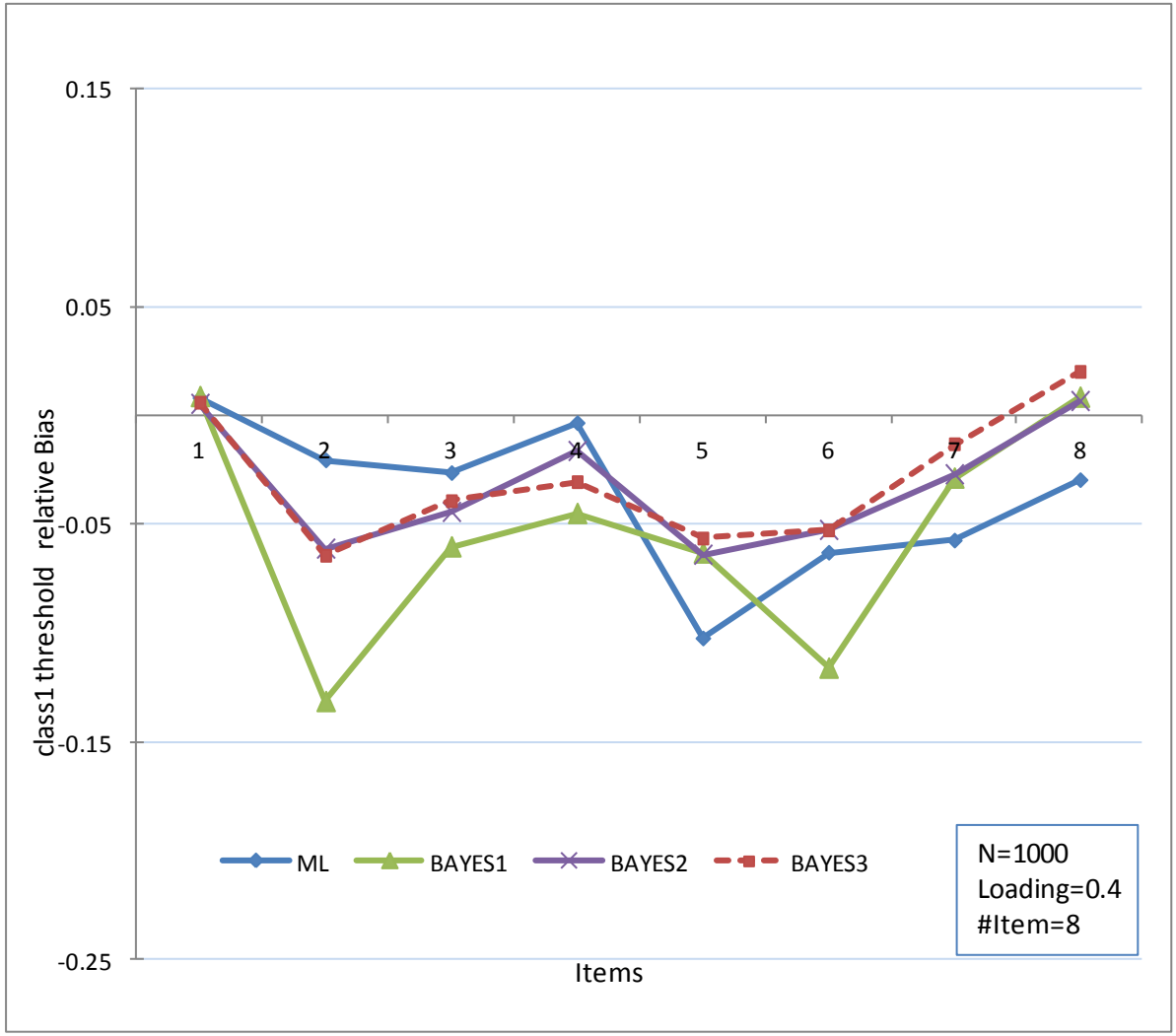


Figure B-15. Relative bias of class 1 threshold for each item when N = 2000, loading = 0.4 and number of items = 8

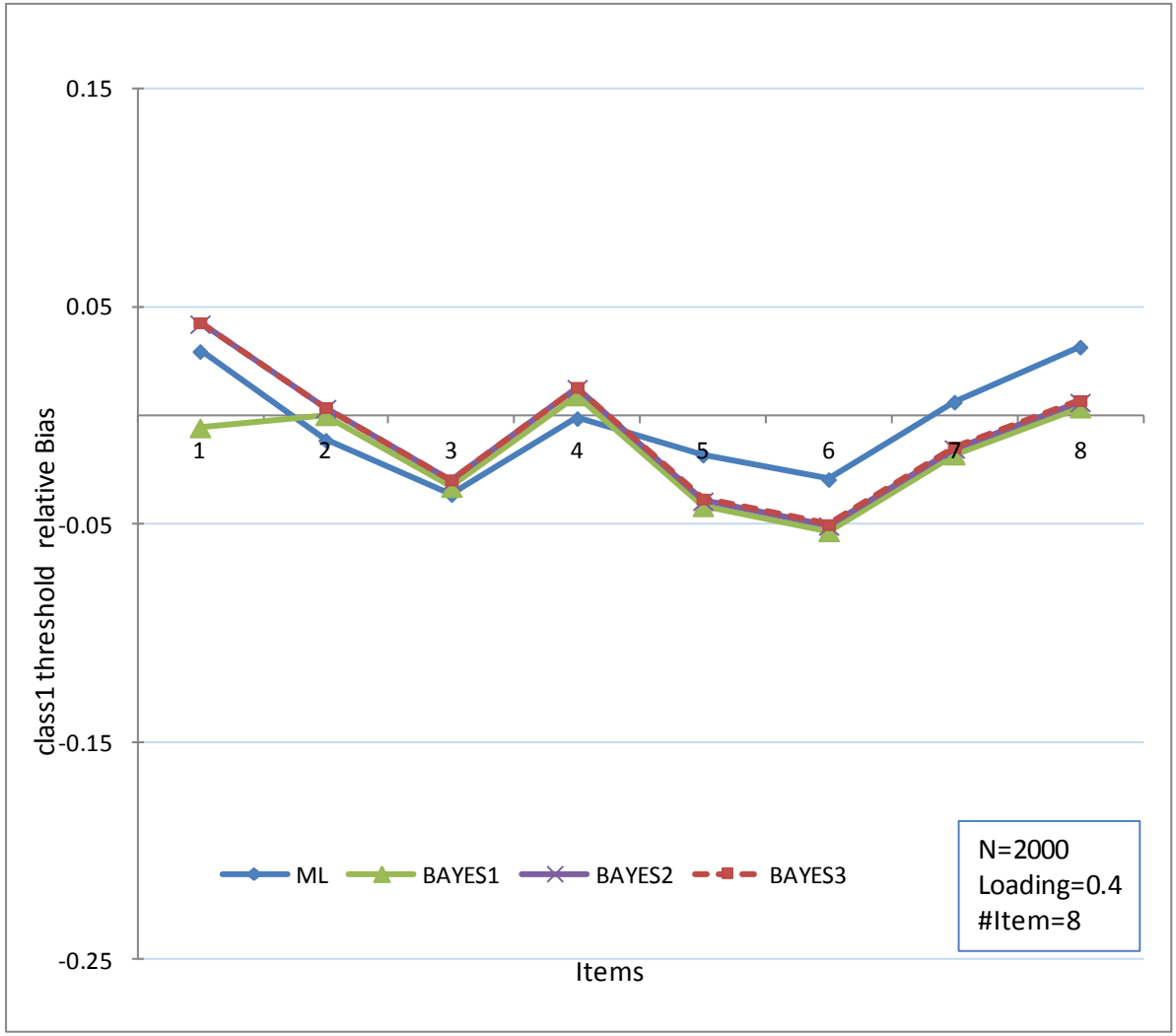


Figure B-16. Relative bias of class 1 threshold for each item when N = 5000, loading = 0.4 and number of items = 8

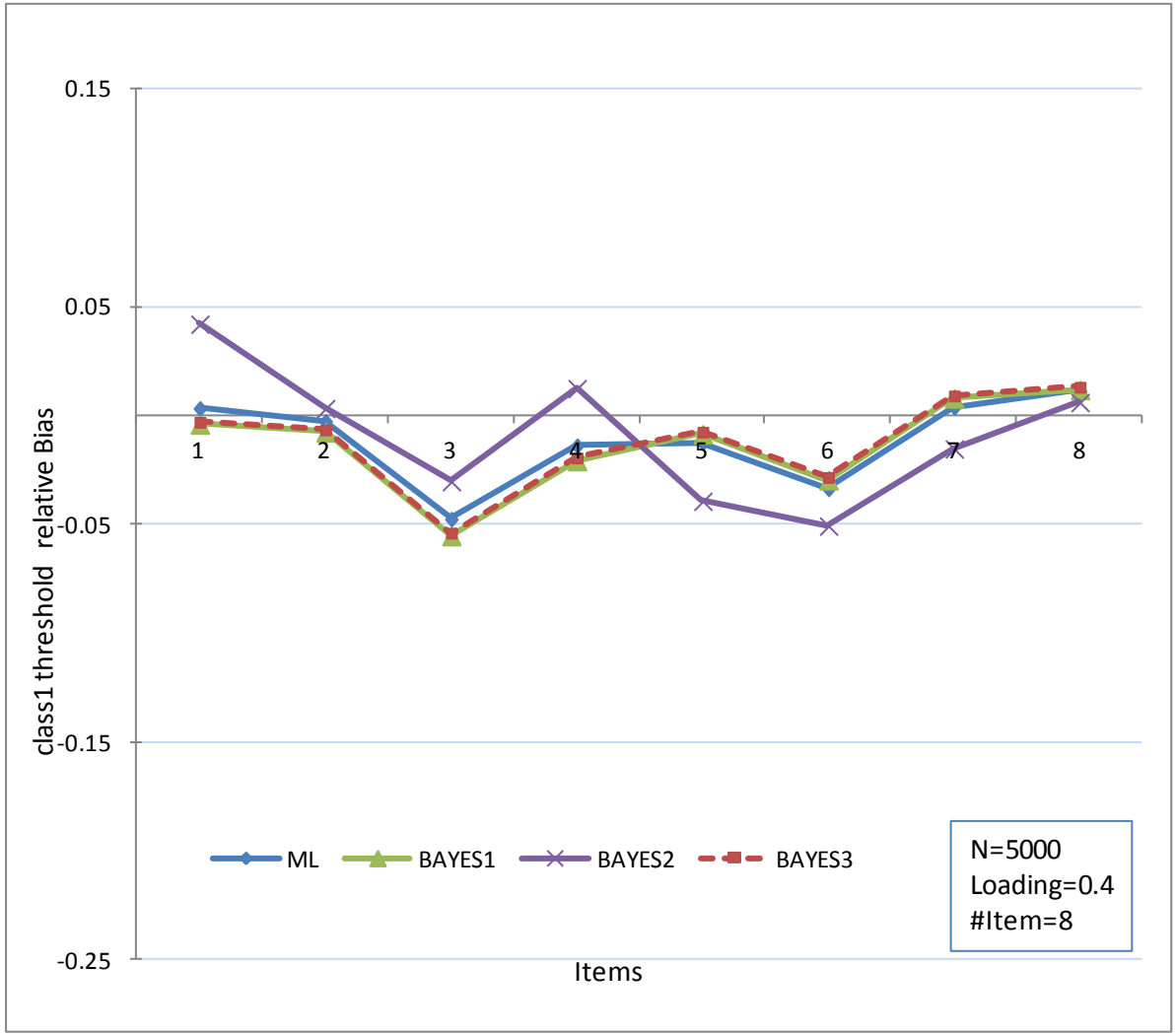


Figure B-17. Relative bias of class 2 threshold for each item when N = 500, loading = 0.8 and number of items = 8

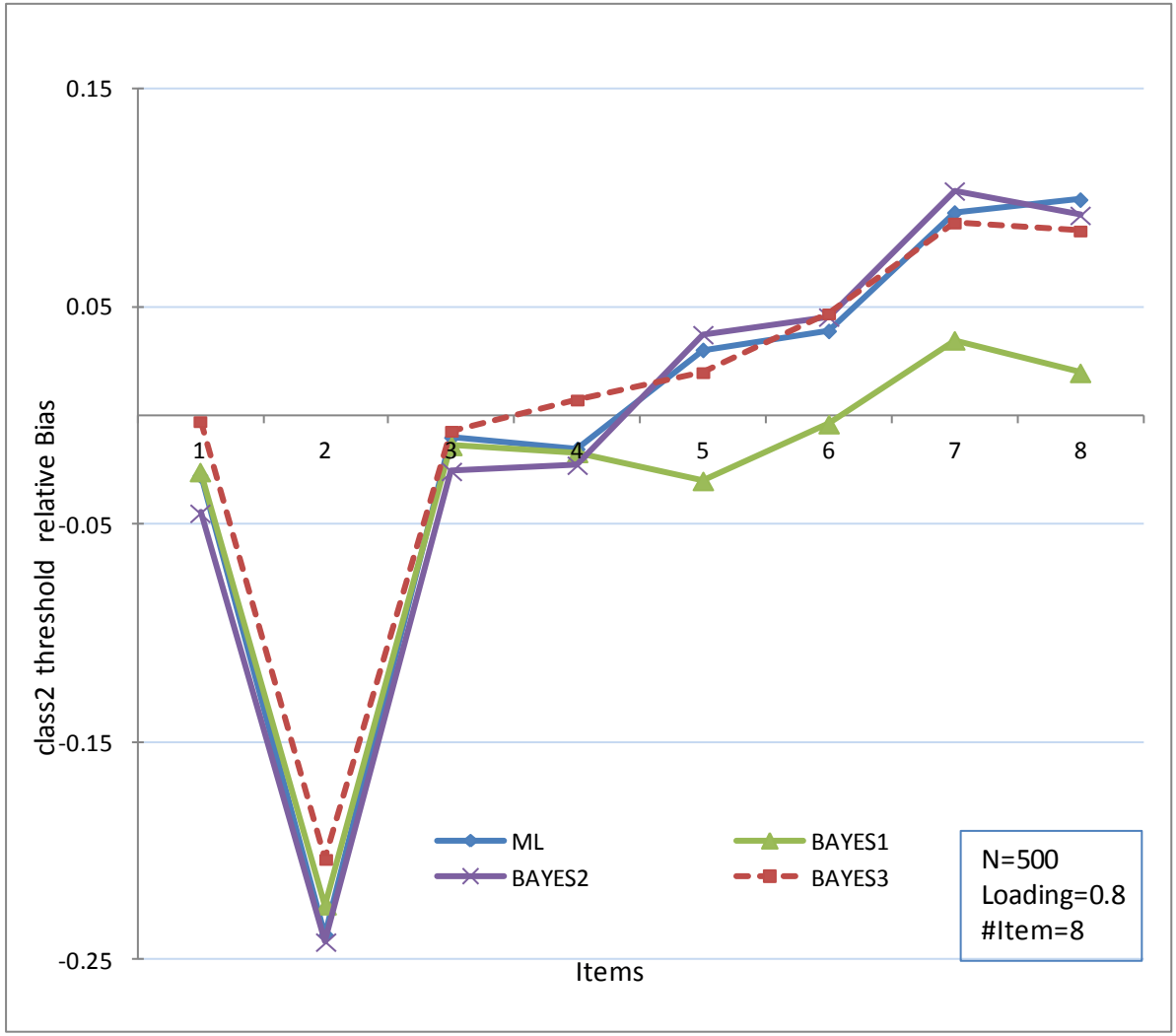


Figure B-18. Relative bias of class 2 threshold for each item when N = 1000, loading = 0.8 and number of items = 8

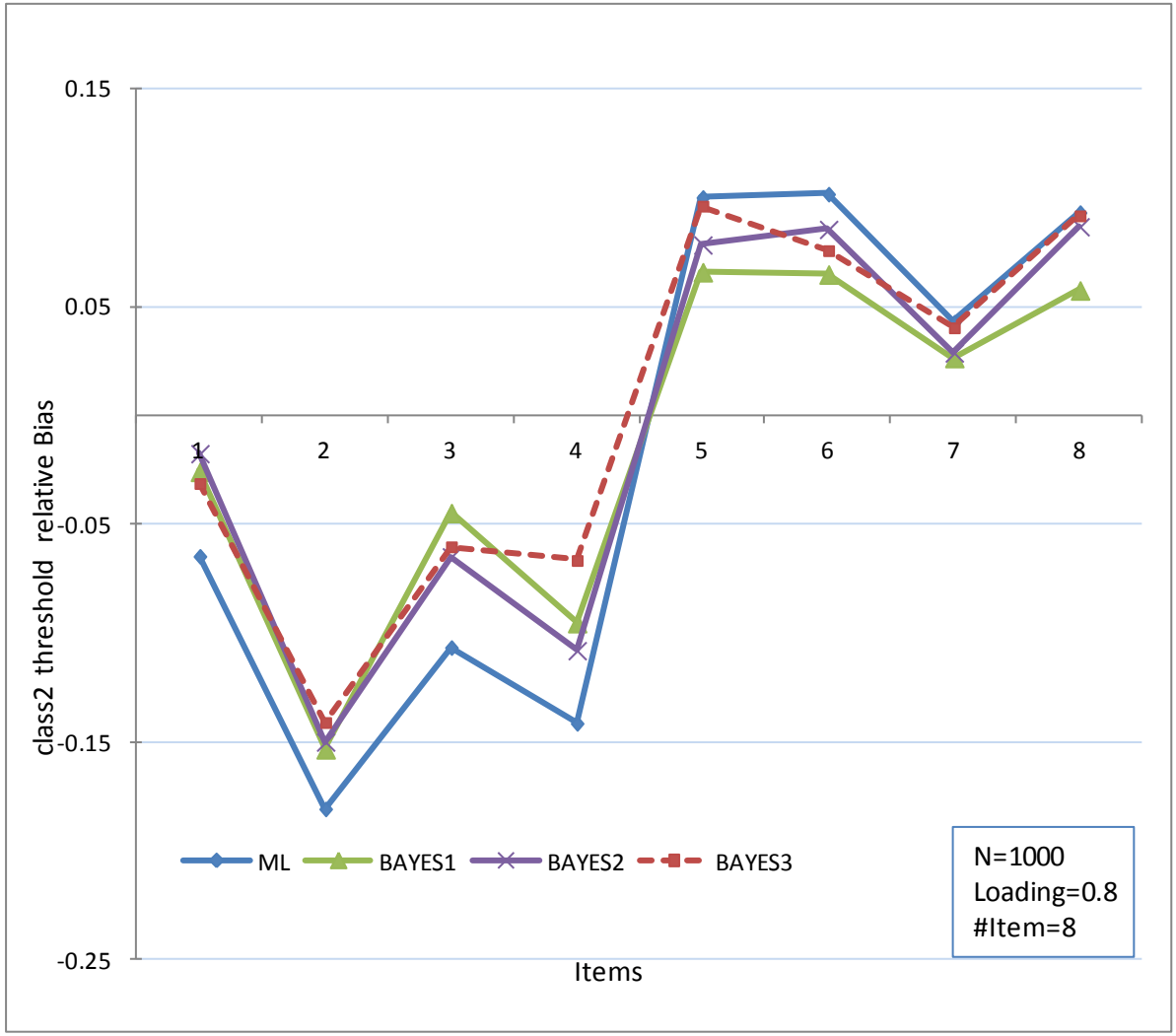


Figure B-19. Relative bias of class 2 threshold for each item when N = 2000, loading = 0.8 and number of items = 8

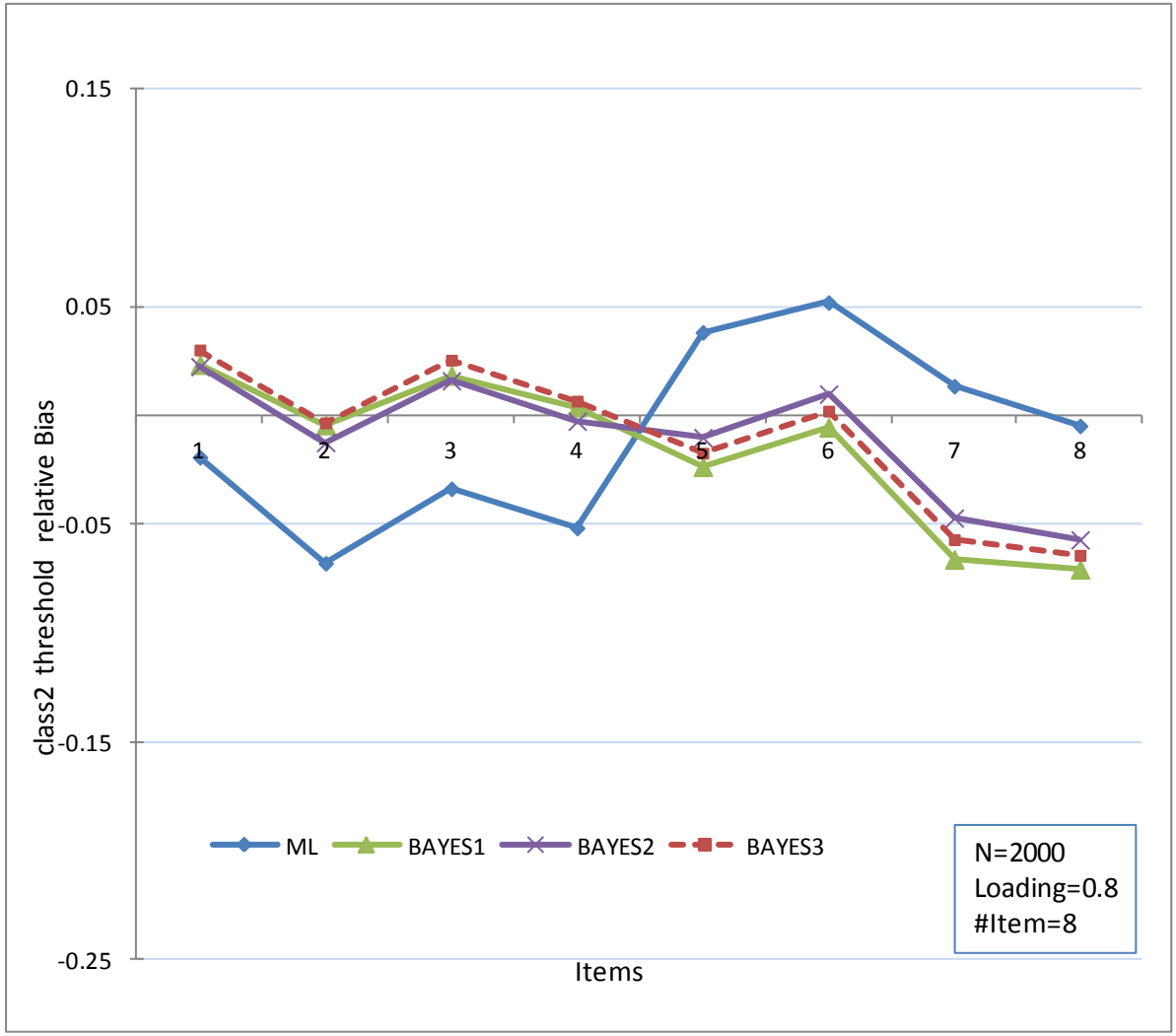


Figure B-20. Relative bias of class 2 threshold for each item when N = 5000, loading = 0.8 and number of items = 8

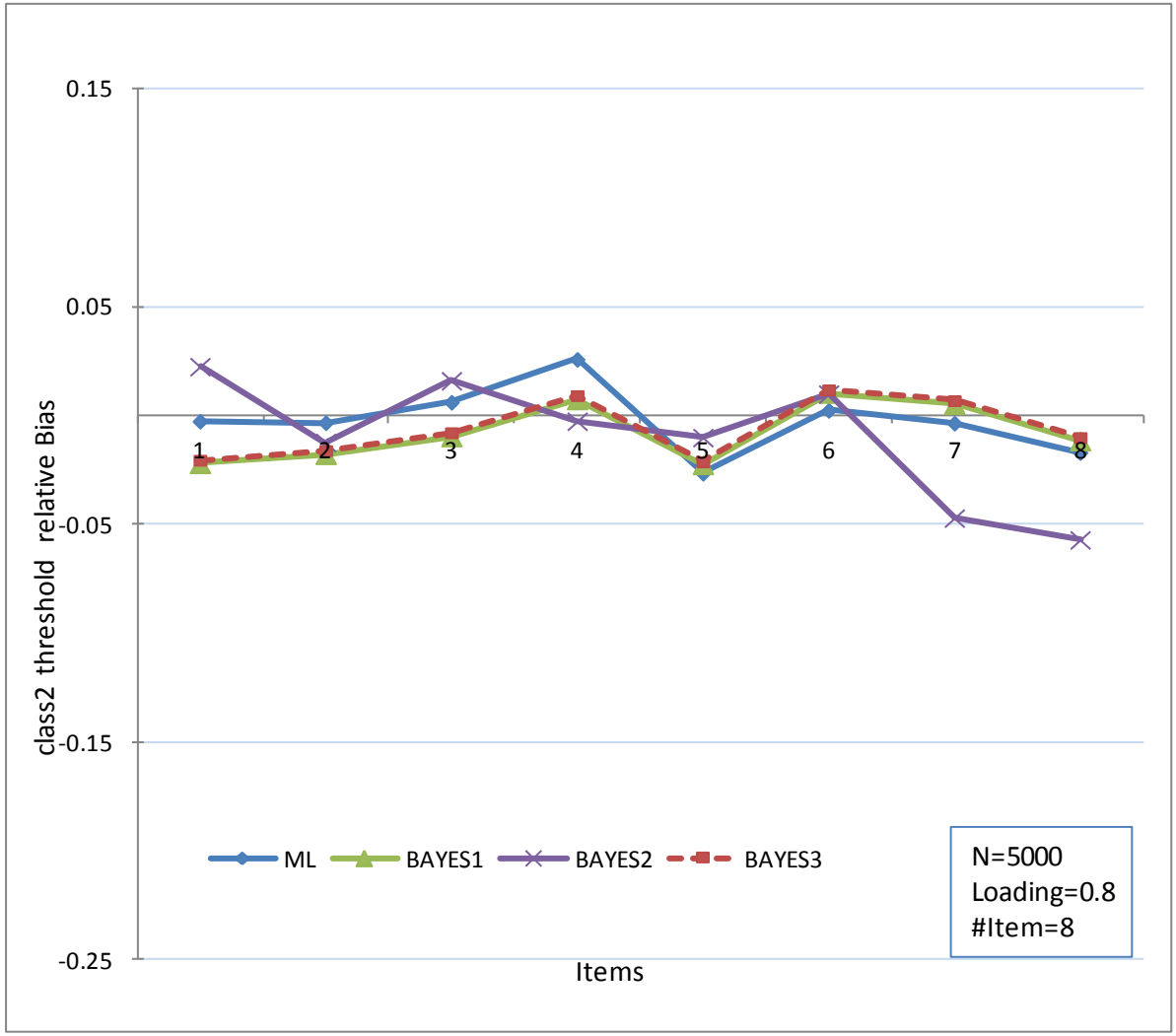


Figure B-21. Relative bias of class 2 threshold for each item when N = 500, loading = 0.4 and number of items = 8

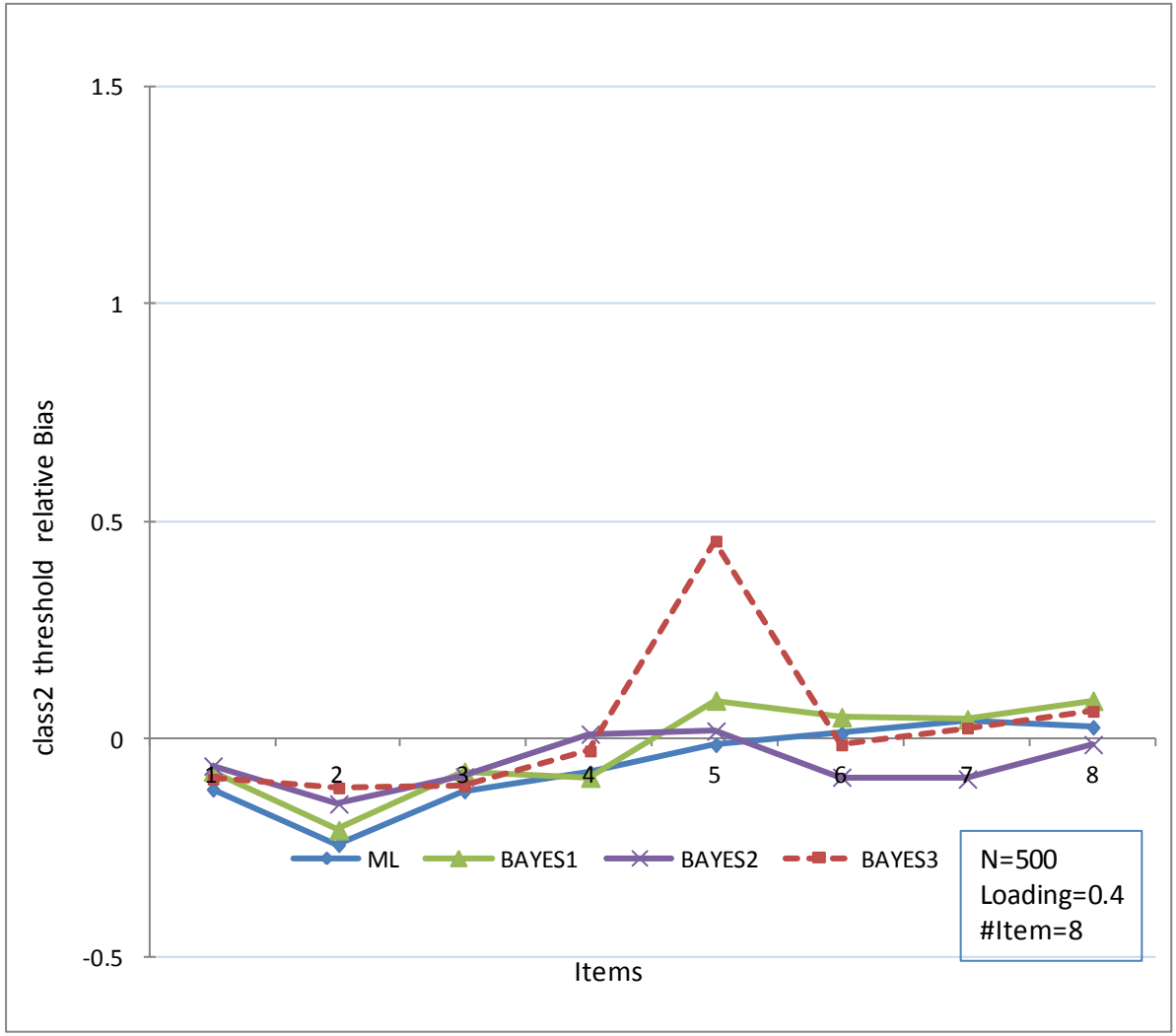


Figure B-22. Relative bias of class 2 threshold for each item when N = 1000, loading = 0.4 and number of items = 8

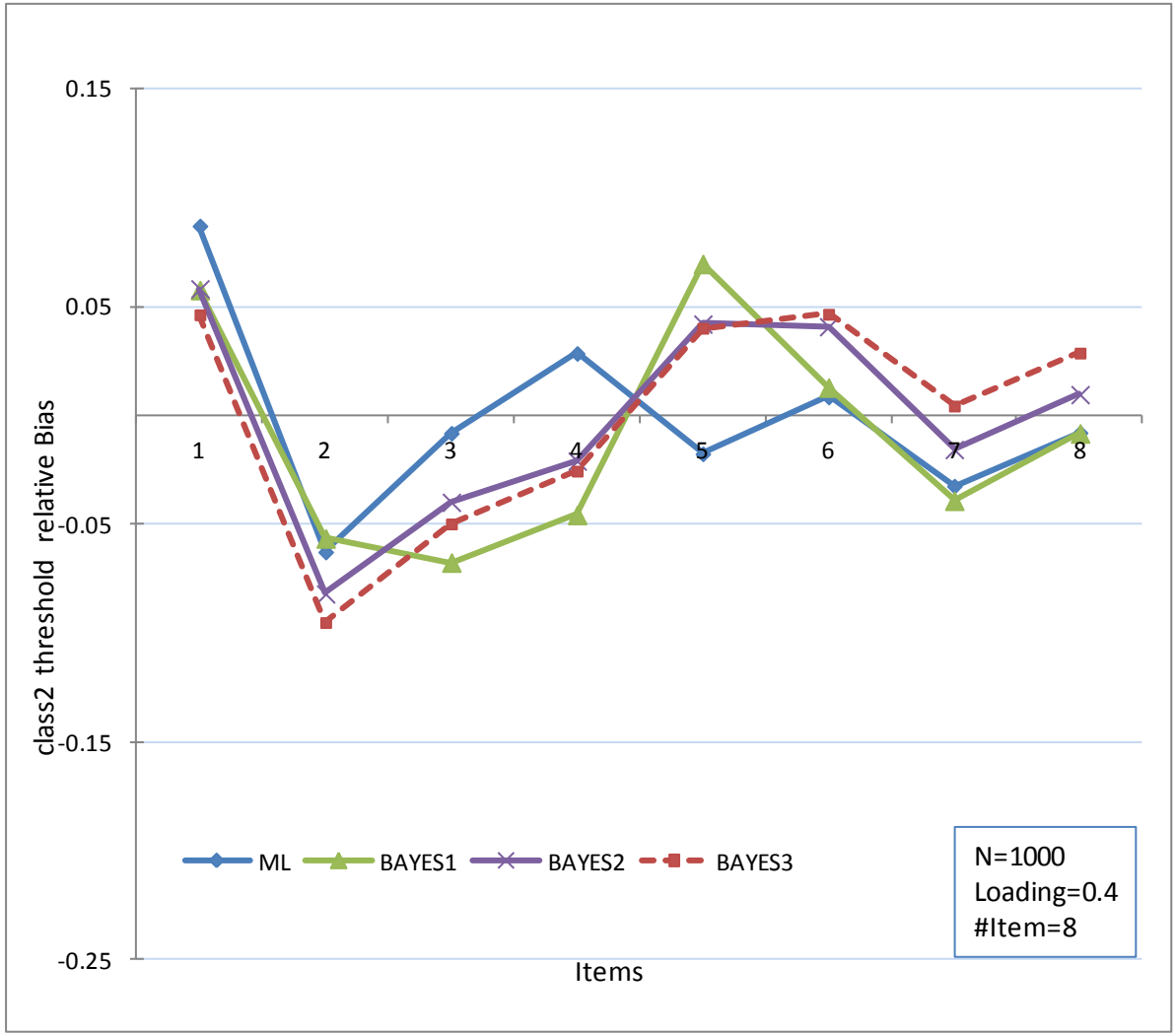


Figure B-23. Relative bias of class 2 threshold for each item when N = 2000, loading = 0.4 and number of items = 8

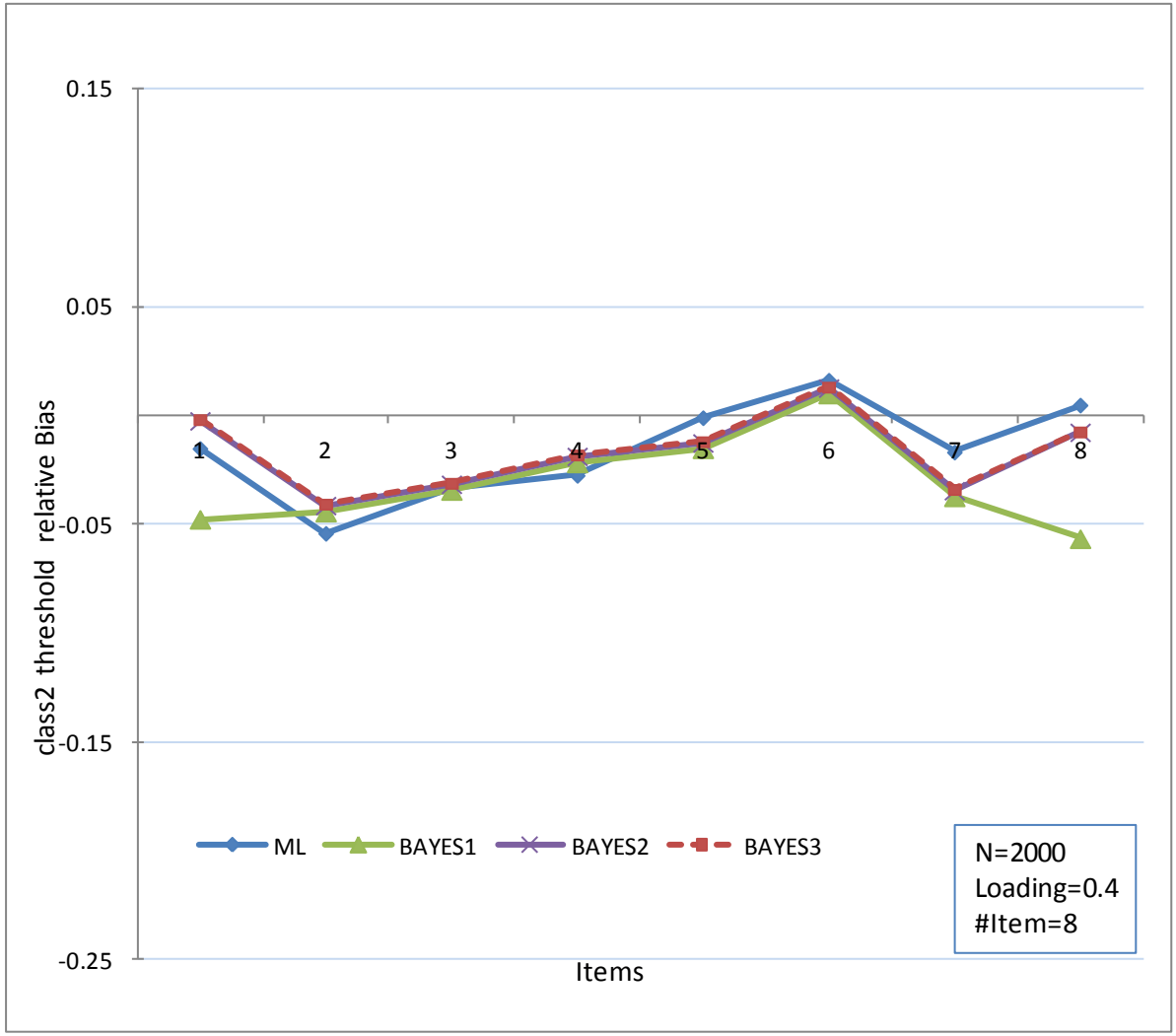


Figure B-24. Relative bias of class 2 threshold for each item when N = 5000, loading = 0.4 and number of items = 8

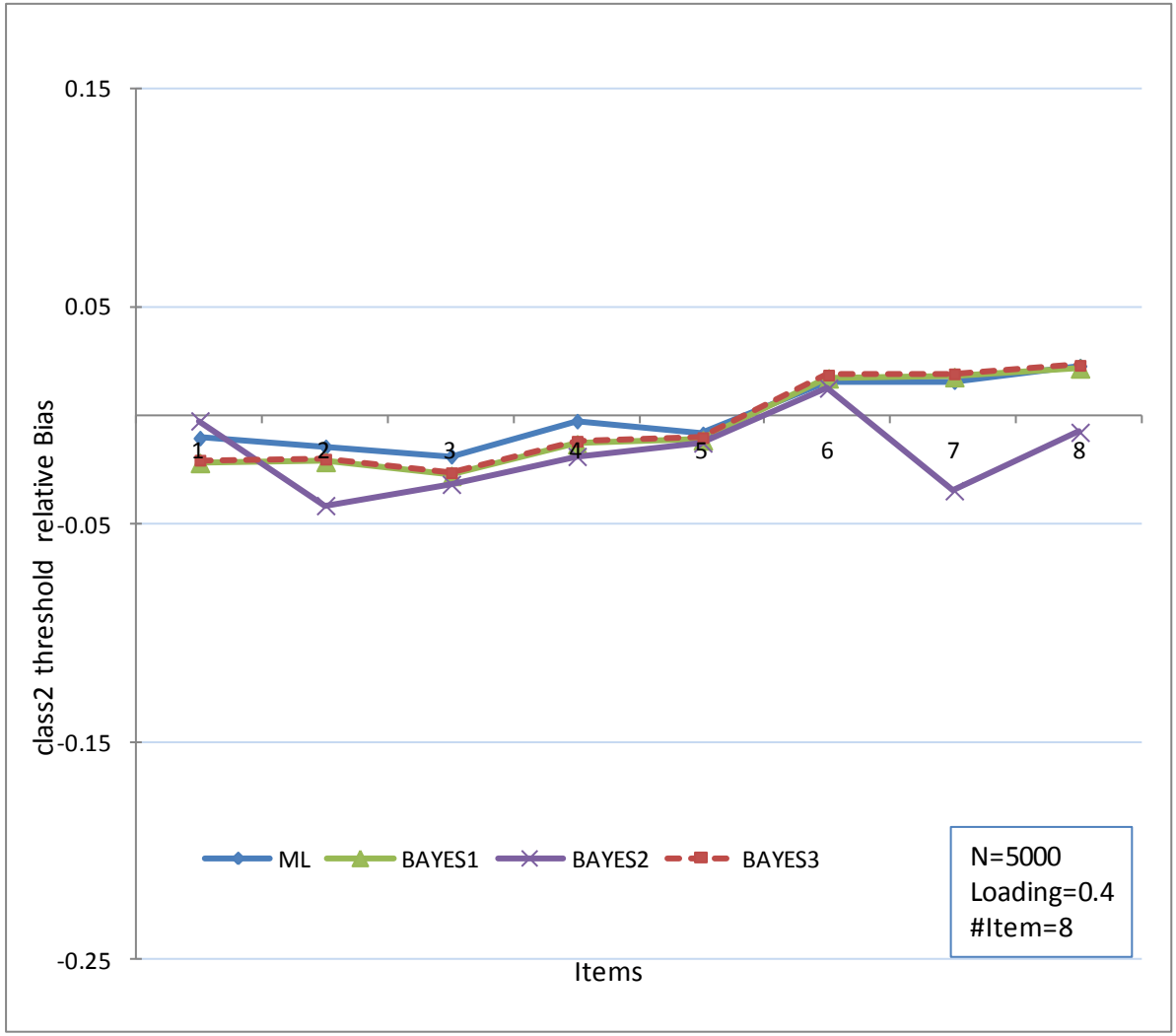


Figure B-25. Relative bias of loading for each item when N = 500, loading = 0.4 and number of items = 30

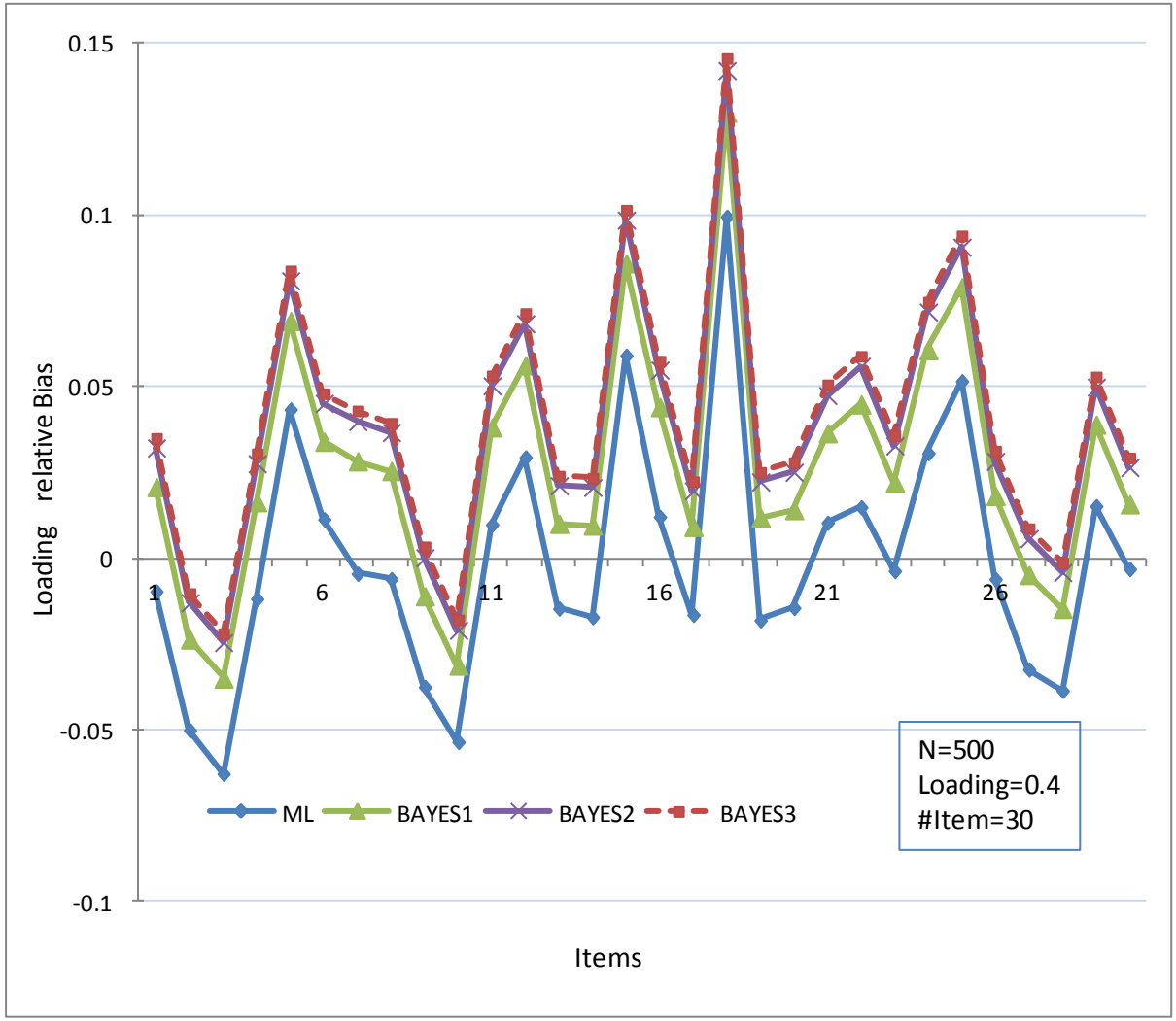


Figure B-26. Relative bias of loading for each item when N = 1000, loading = 0.4 and number of items = 30

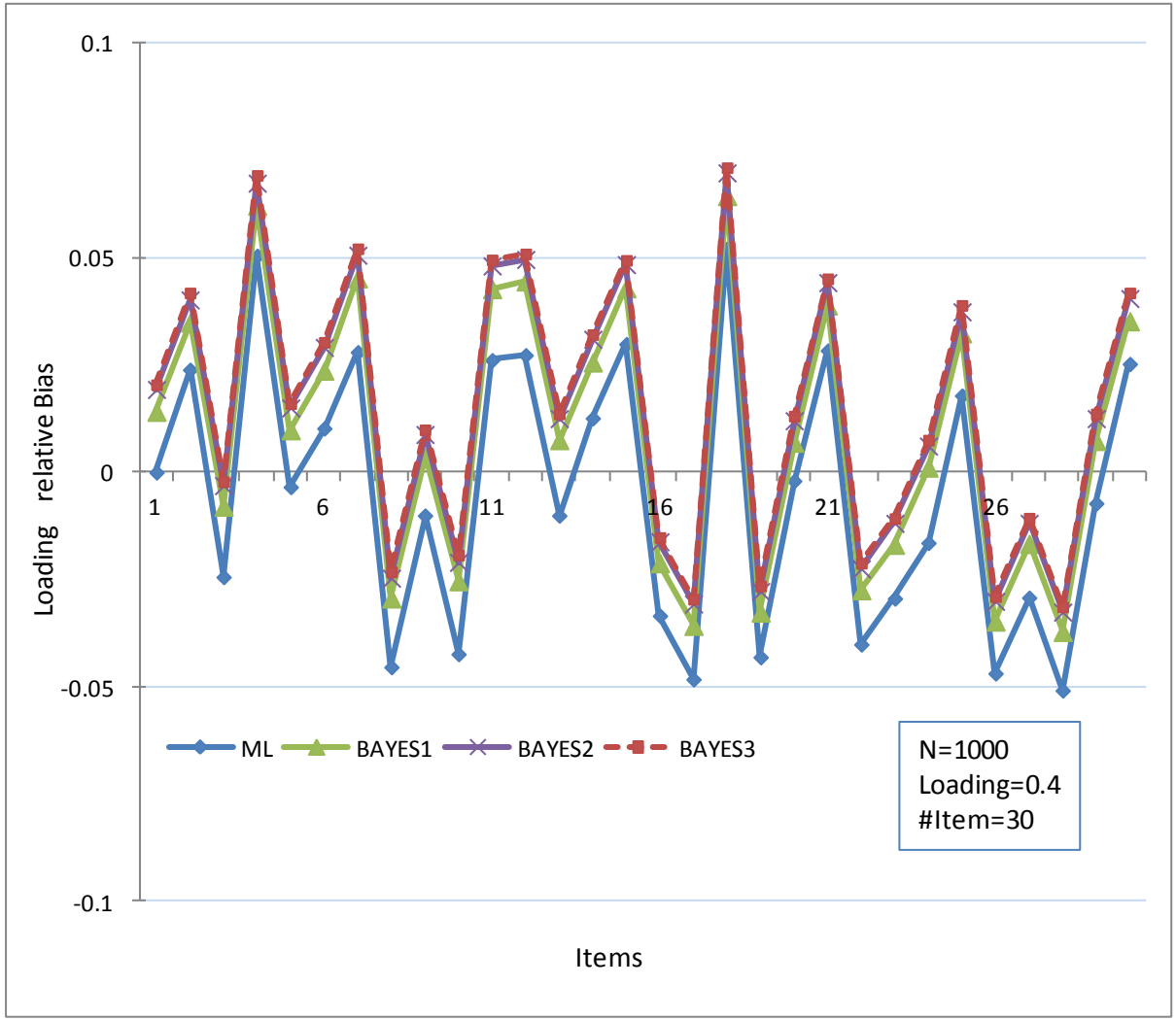


Figure B-27. Relative bias of loading for each item when N = 2000, loading = 0.4 and number of items = 30

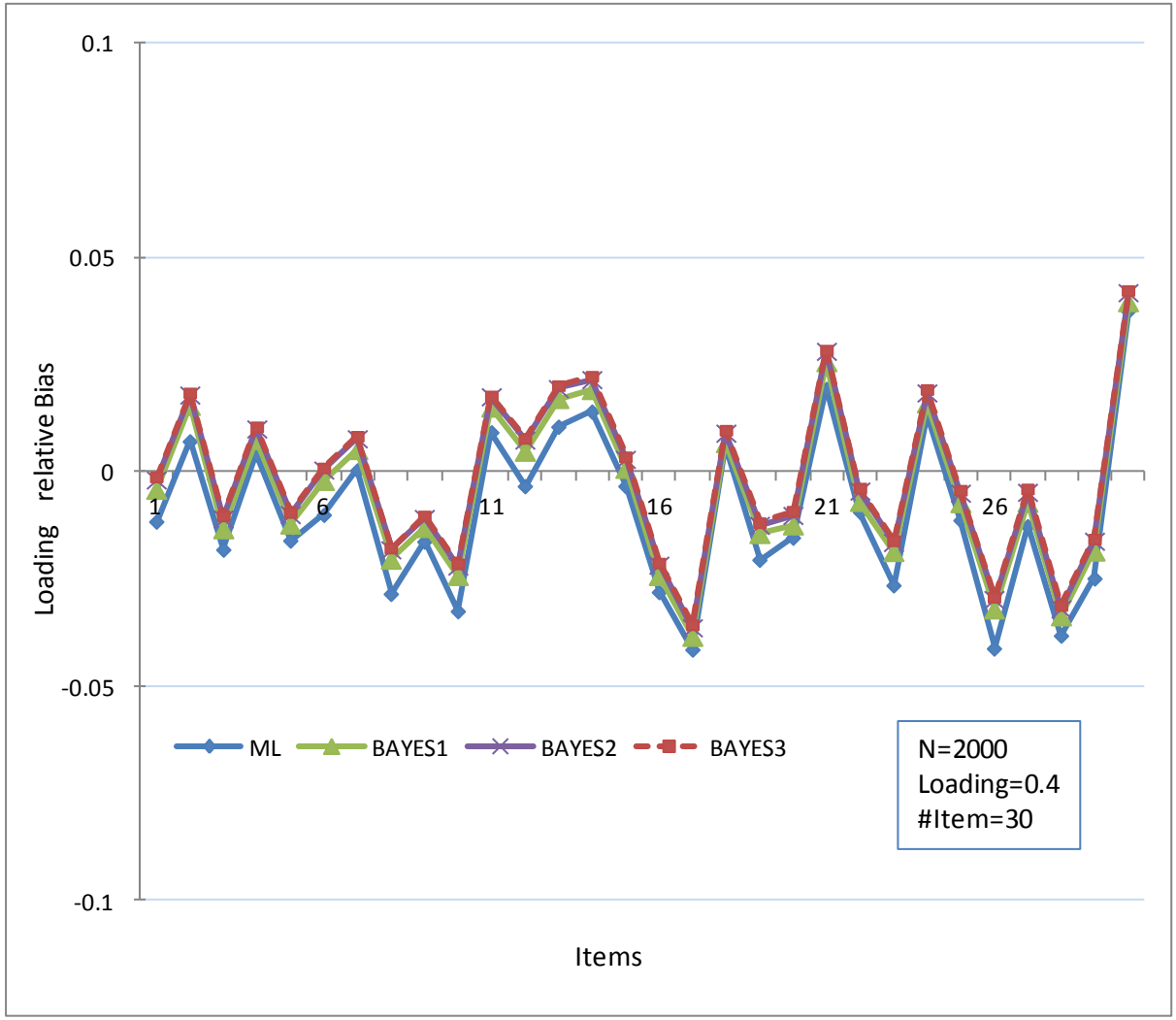


Figure B-28. Relative bias of loading for each item when N = 5000, loading = 0.4 and number of items = 30

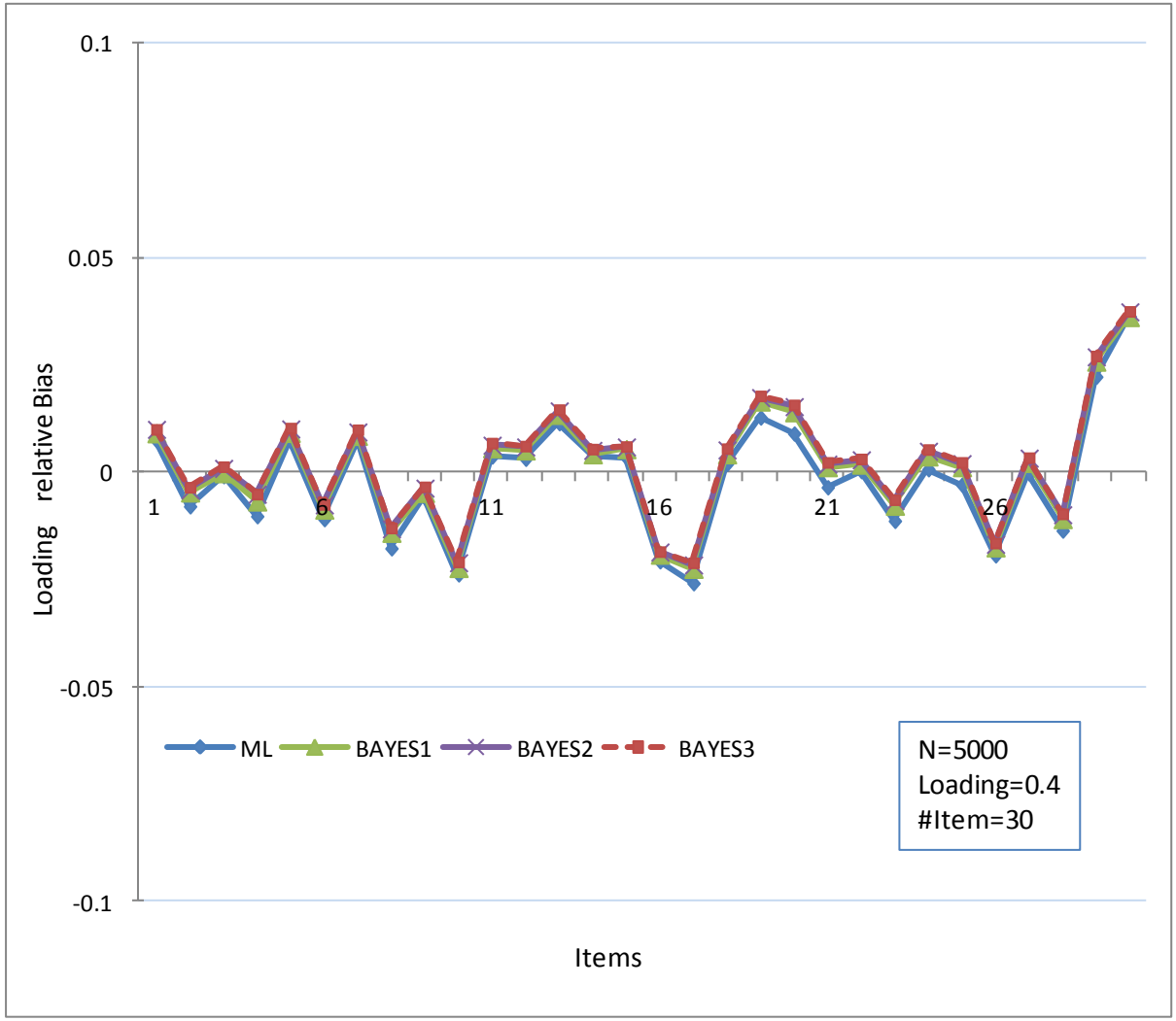


Figure B-29. Relative bias of class 1 threshold for each item when N = 500, loading = 0.4 and number of items = 30

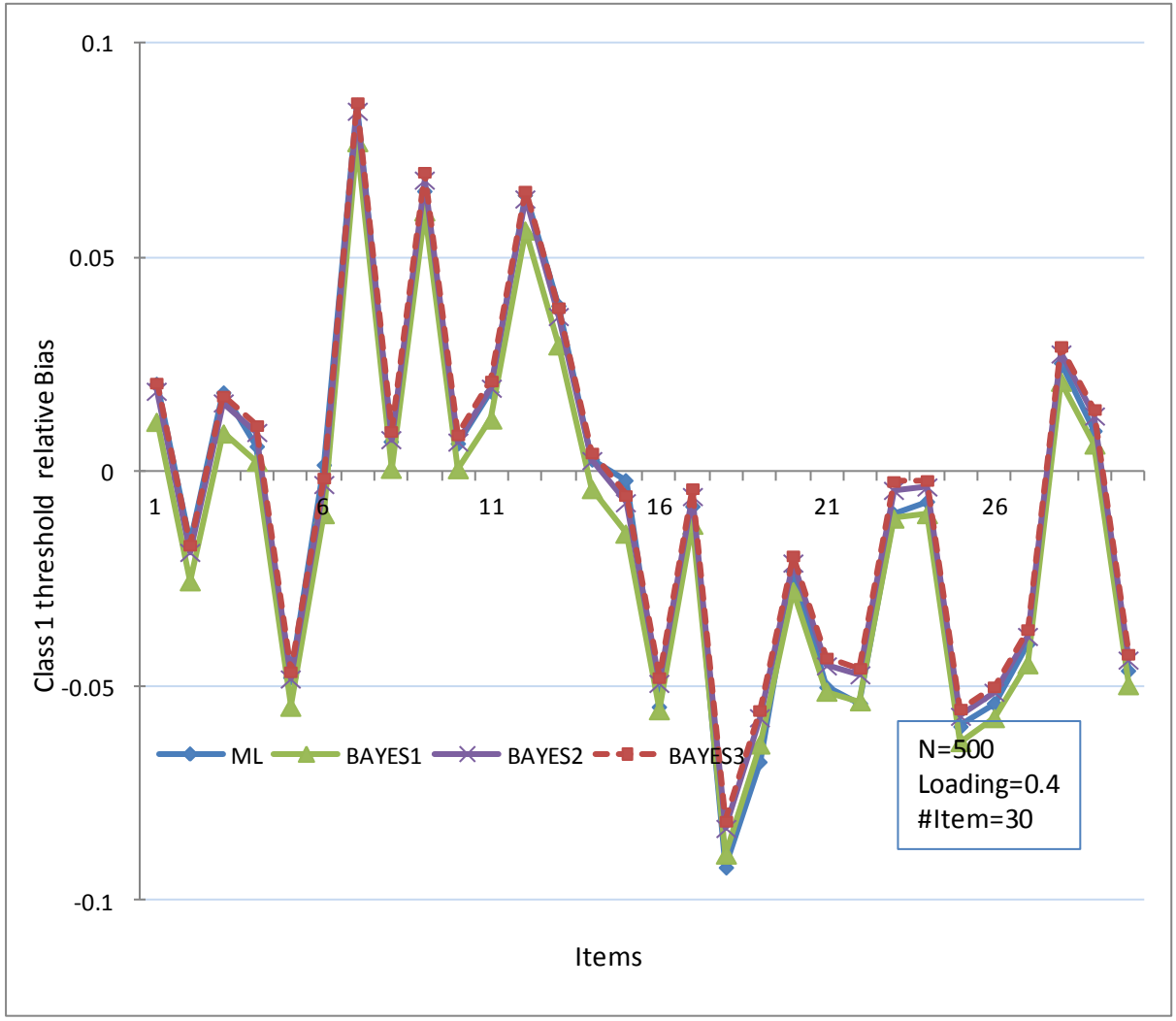


Figure B-30. Relative bias of class 1 threshold for each item when N = 1000, loading = 0.4 and number of items = 30

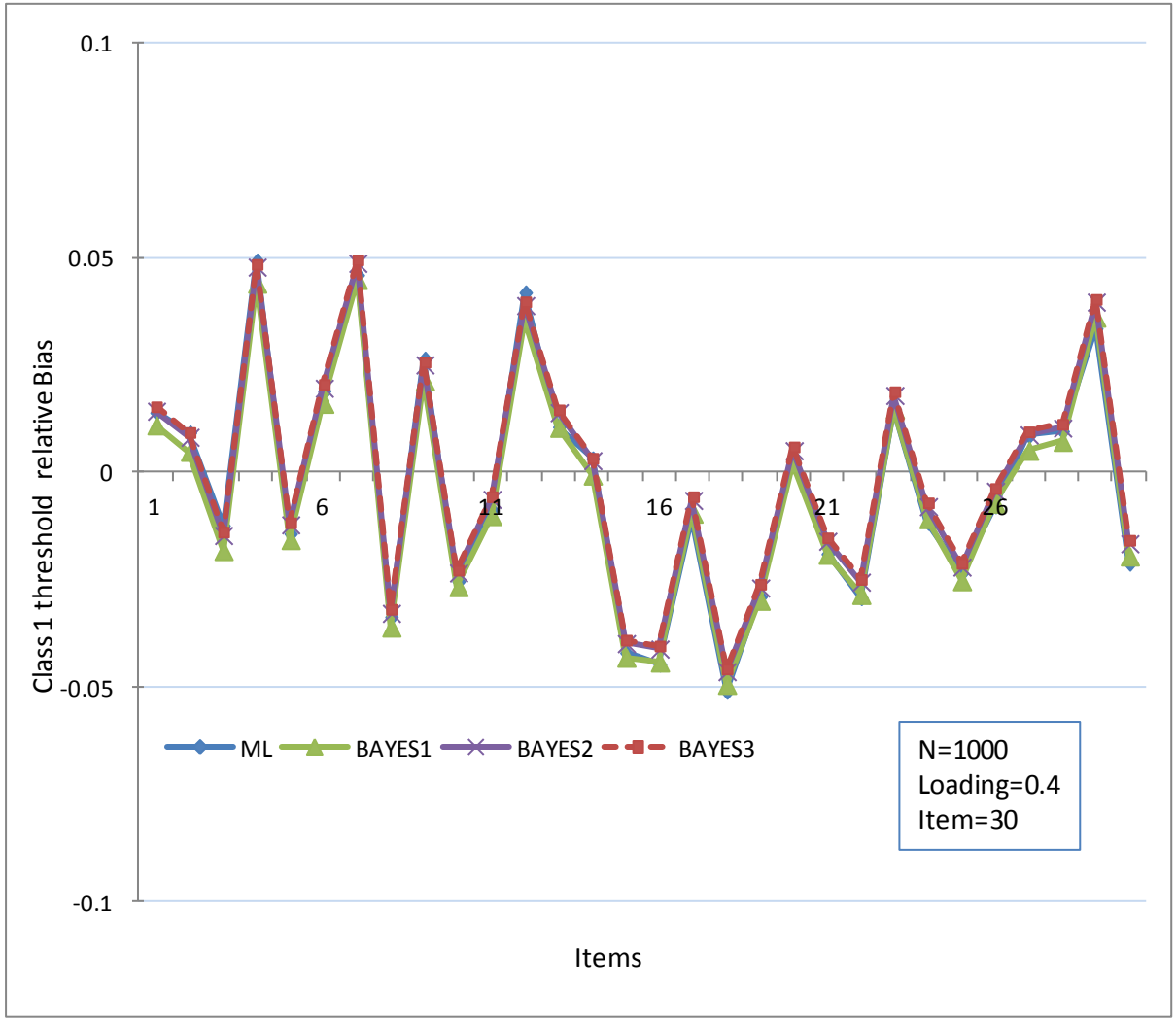


Figure B-31. Relative bias of class 1 threshold for each item when N = 2000, loading = 0.4 and number of items = 30

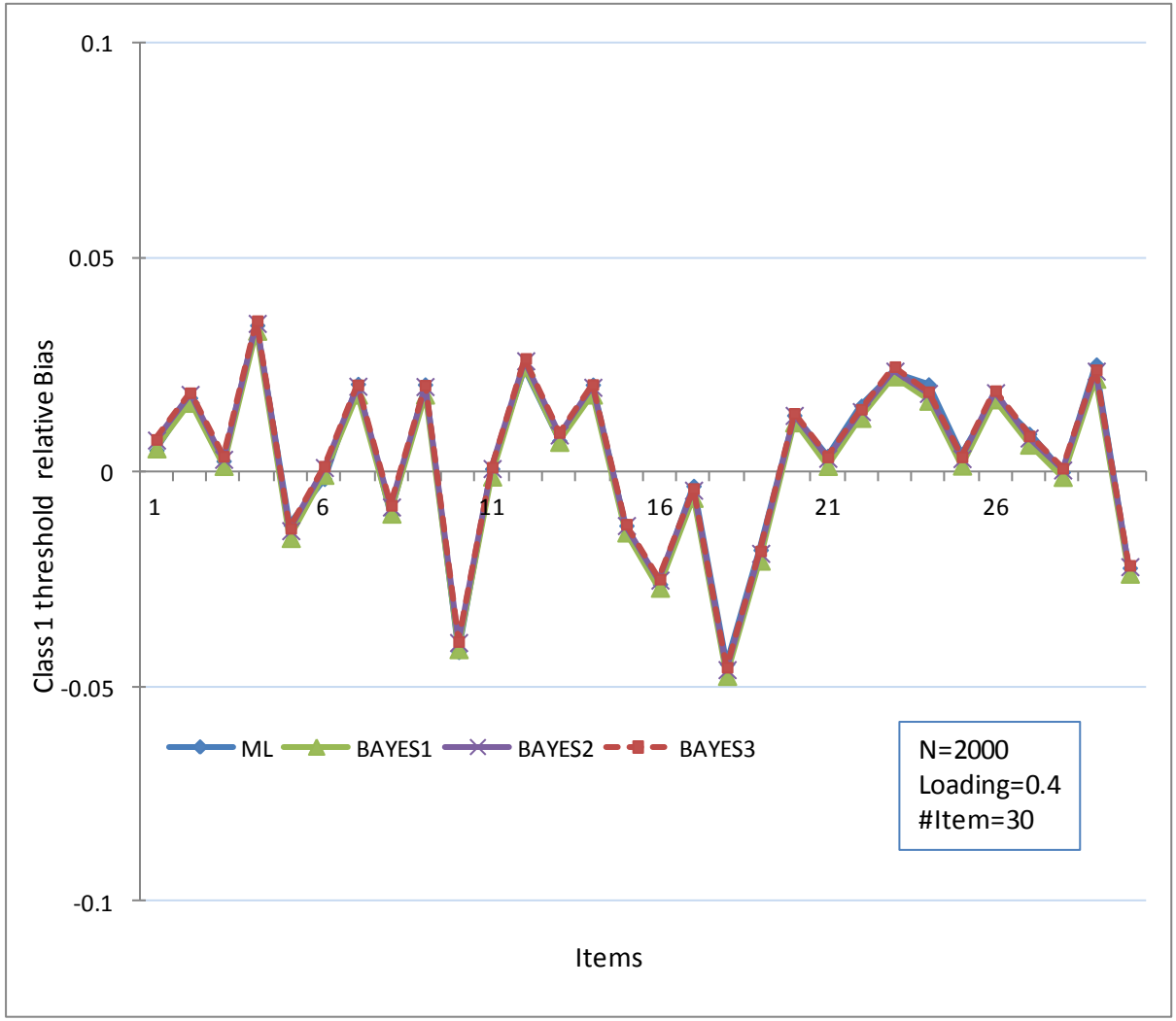


Figure B-32. Relative bias of class 1 threshold for each item when N = 5000, loading = 0.4 and number of items = 30

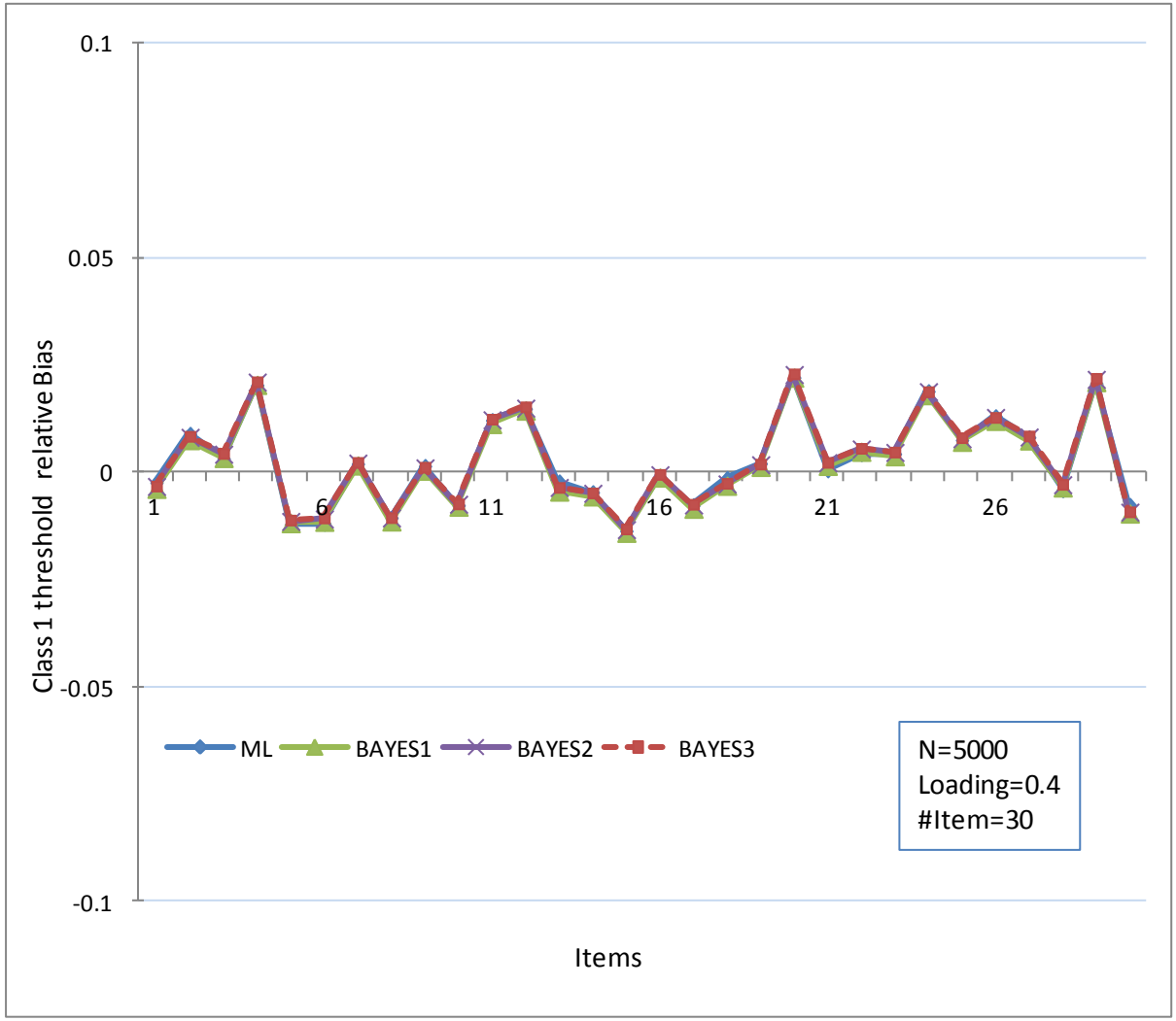


Figure B-33. Relative bias of class 2 threshold for each item when N = 500, loading = 0.4 and number of items = 30

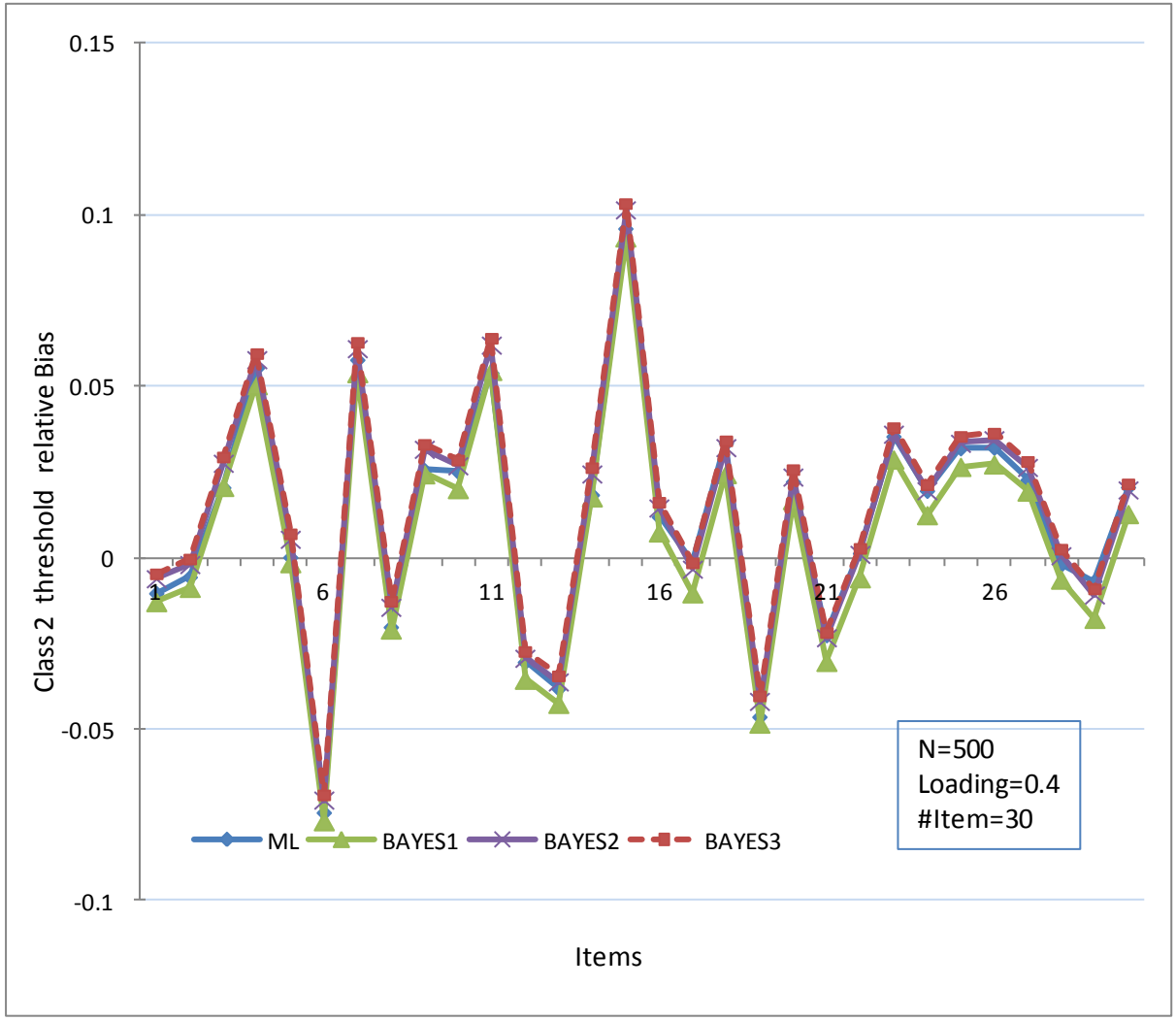


Figure B-34. Relative bias of class 2 threshold for each item when N = 1000, loading = 0.4 and number of items = 30

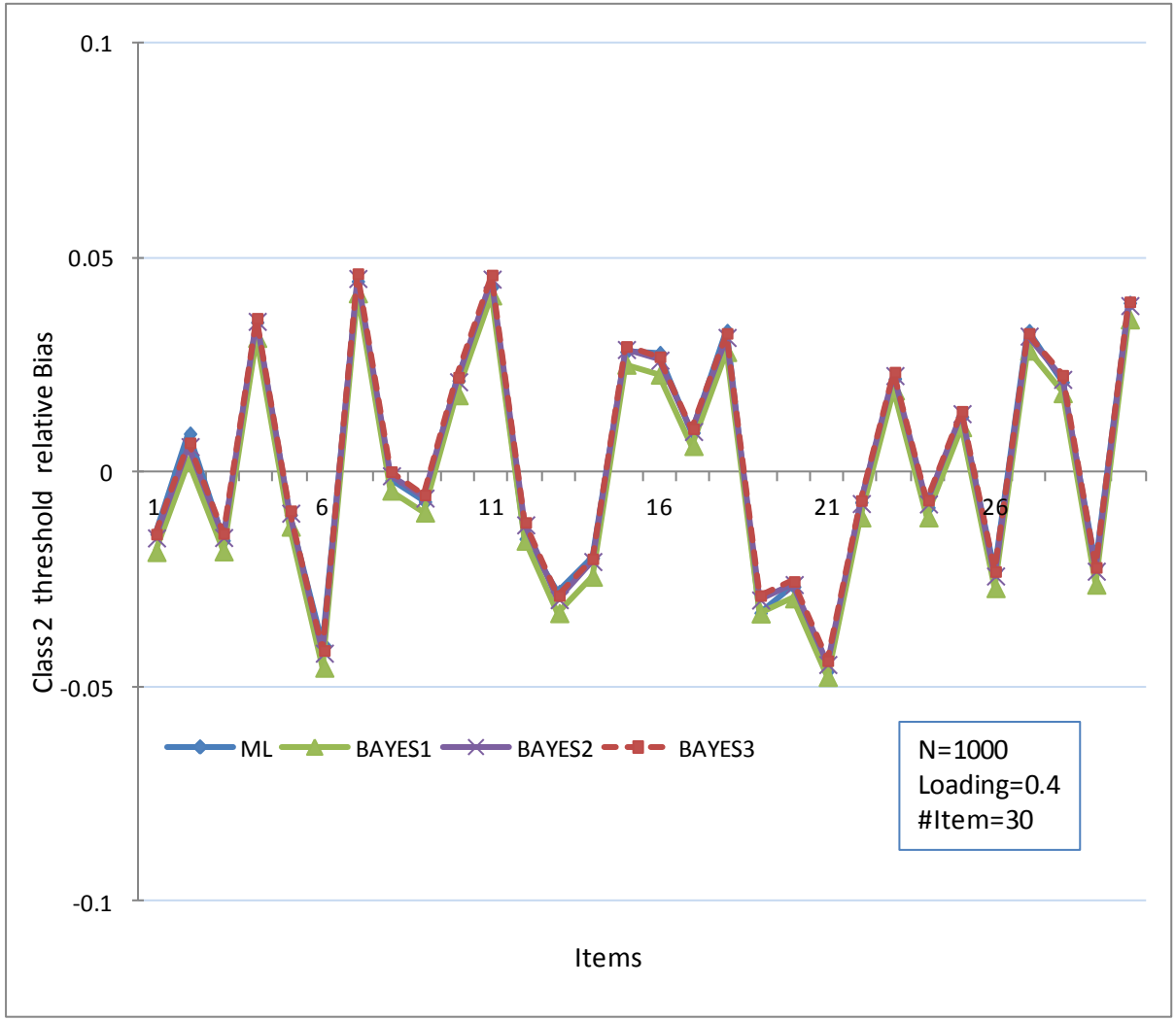


Figure B-35. Relative bias of class 2 threshold for each item when N = 2000, loading = 0.4 and number of items = 30

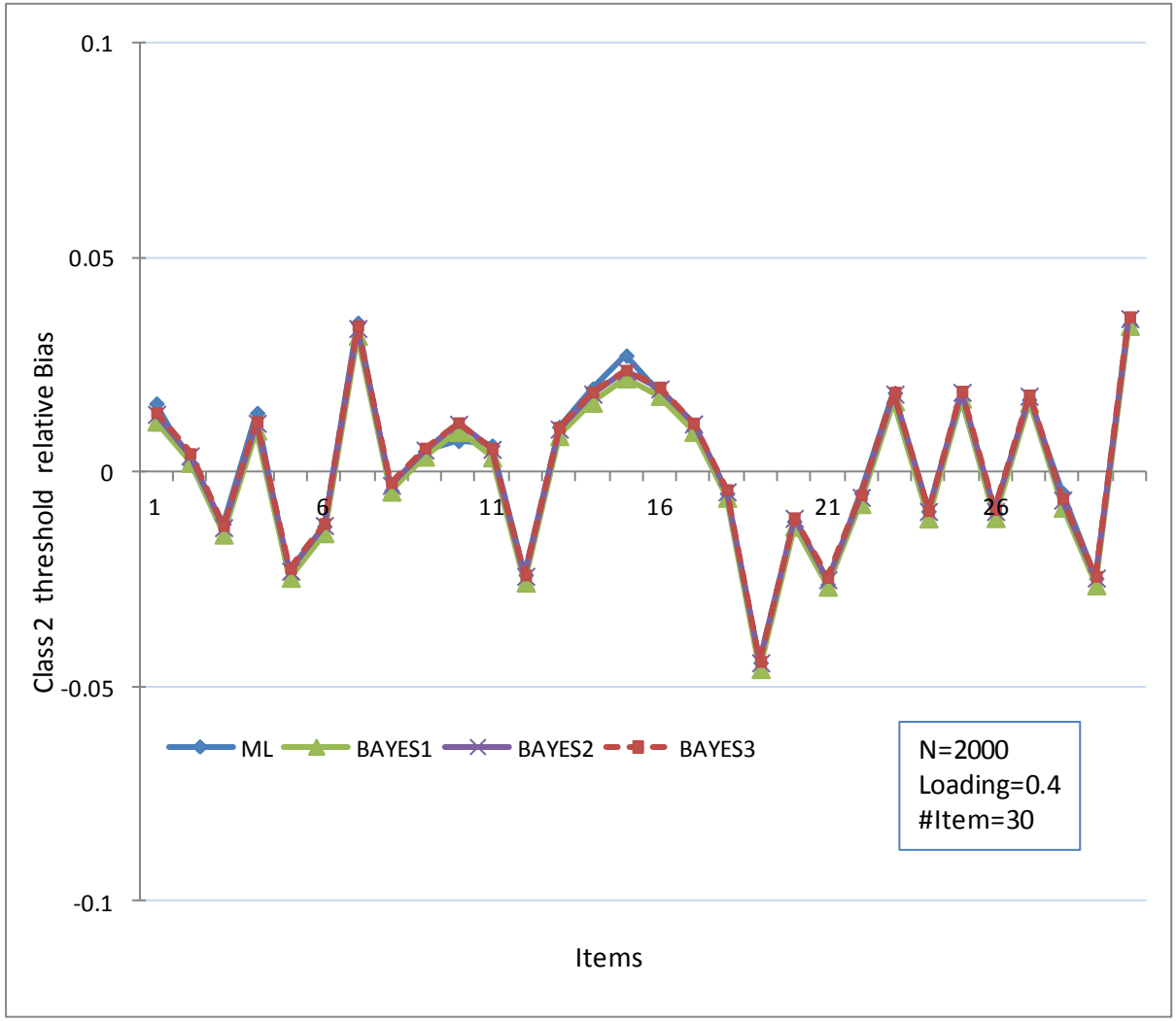
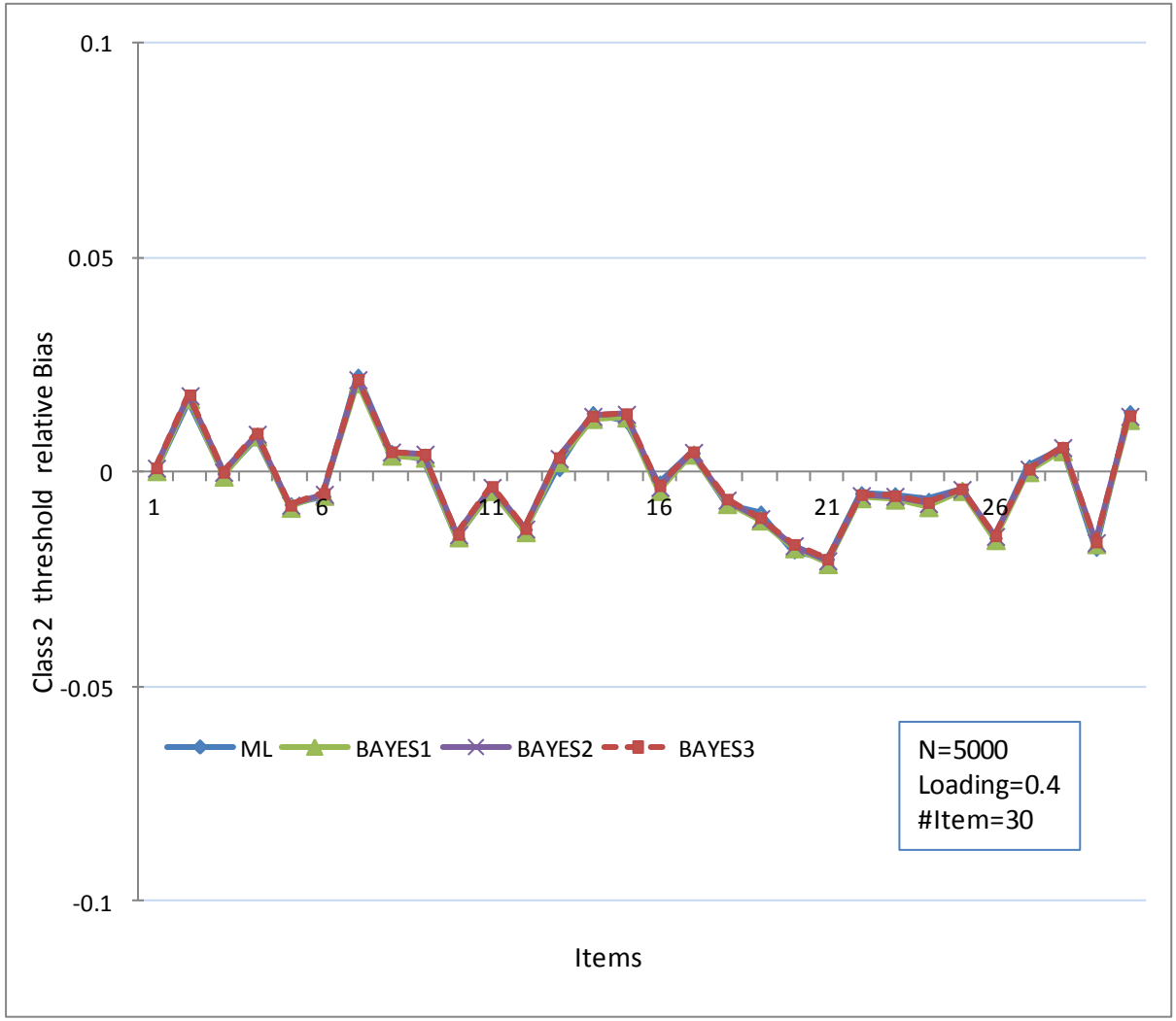


Figure B-36. Relative bias of class 2 threshold for each item when N = 5000, loading = 0.4 and number of items = 30



References

Asparouhov, T., & Muthén, B. (2010a). *Bayesian analysis using Mplus* (Technical Report). Los Angeles, CA: Muthén & Muthén.

Asparouhov, T., & Muthén, B. (2010b). *Bayesian analysis of latent variable models using Mplus.* (Technical report). Los Angeles, CA: Muthén & Muthén.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational & Behavioral Statistics, 26*, 381-409.

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

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 3,* 473-514.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13,* 195-212.

Clark, S.L. (2010). *Mixture modeling with behavioral data.* (Doctoral dissertation, University of California, Los Angeles). Retrieved from http://www.statmodel.com/papers.shtml

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

Cohen, A. S., Gregg, N., & Deng, M. (2005, April). *A mixture IRT model for testlets*. Paper presented at the annual meeting of the American Educational Research Association, Montréal, Canada.

Dai, Y. (2009). *A mixture Rasch model with a covariate: A simulation study via Bayesian Markov Chain Monte Carlo estimation*. (Doctoral dissertation, University of Maryland). Available from ProQuest Dissertations and theses database. (UMI No. 3391214)

Dayton, C. M., & Macready, G. B. (2007). Latent class analysis in psychometrics. In C. R. Rao & S. Sinharay (Eds), *Handbook Of Statistics* (pp. 421-446), North-Holland, The Netherland: Elsevier.

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

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

Gagné, P. E. (2004). *General confirmatory factor mixture models: A tool for assessing factorial invariance across unspecified populations.* (Doctoral dissertation, University of Maryland). Available from ProQuest Dissertations and theses database. (UMI No. 3125434)

Gagné, P. E. (2006). Mean and covariance structure mixture models. In G. R. Hancock, & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 197-224). Greenwich, CT: Information Age Publishing.

Gagné, P. E., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research, 41,* 65-83.

Gamerman, D. (1997). *Markov Chain Monte Carlo.* New York, NY: Chapman & Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Gelman A., Jakulin A., Pittau M. G., & Su Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics, 2,* 1360-1383.

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

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

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

Green, B. F. (1951). A general solution for the latent class model of latent structure analysis. *Psychometrika, 16,* 151-160.

Green, B. F. (1952). Latent structure analysis and its relation to factor analysis. *Journal of the American Statistical Association, 47,* 71-76.

Grosuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hancock, G. R. (2004). Experimental, quasi-experimental, and nonexperimental design and analysis with latent variables. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 317-334). Thousand Oaks, CA: Sage.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and manova in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling: A Multidisciplinary Journal, 7*, 534-556.

Hancock G. R., & Mueller. R. O. (2006). *Structural equation modeling: A second course.* Greenwich, CT: Information Age Publishing.

Hancock, G. R., & Samuelsen, K. (2008). *Mixture models in latent variable research*. Greenwich, CT: Information Age Publishing.

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

Hou, X., & Hancock, G. R. (2010, May). *Model selection methods for factor mixture model with dichotomous outcomes.* Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Jiao, H., von Davier, M., & Wang, S. (2010a, May). *Polytomous mixture Rasch testlet model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Jiao, H., von Davier, M., & Wang, S. (2010b, May). *Marginal maximum likelihood estimation of the Rasch mixture testlet model.* Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183-202.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15,* 136-153.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association. 91,* 1343–1370.

Kelderman, H. & Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27,* 307-327.

Kim, Y. K., & Muthén, B. (2009). Two-part factor mixture modeling: Application to an aggressive behavior measurement instrument. *Structural Equation Modeling, 16,* 602-624.

Lange, K. (1995). A quasi-Newton acceleration on the EM algorithm. *Statistica Sinica, 5,* 1-18.

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

Lee, S. Y. (2007). *Structural equation modeling. A Bayesian approach.* New York, NY: John Wiley & Sons.

Lee, S. Y., Song X. Y., & Cai, J. H. (2010). A Bayesian approach for nonlinear structural equation models with dichotomous variables Using logit and probit links. *Structural Equation Modeling, 17,* 280-302.

Leite, W., & Cooper, L. (2010). Detecting social desirability bias using factor mixture models. *Multivariate Behavioral Research, 45,* 271-293.

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

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*, 21-39.

Lubke, G. H., & Muthén, B. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling, 14*, 26-47.

Lubke, G., Muthén, B., Moilanen, I., McGough, J., Loo, S., Swanson, J., … Smalley, S. (2007). Subtypes versus severity differences in the attention-deficit/hyperactivity disorder in the northern Finnish birth cohort. *Journal of the American Academy of Child and Adolescent Psychiatry, 46,* 1584-1593.

Lubke, G. H., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research, 10*, 499-532.

Lubke, G. H., & Neale, M. C. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research, 43,* 592-620.

Lubke, G. H., & Spies, J. (2008). Choosing a "correct" factor mixture model: Power, limitations, and graphical data exploration. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 343-362). Charlotte, NC: Information Age Publishing.

Lynch, S. M. (2007*). Introduction to applied Bayesian statistics and estimation for social scientists.* New York, NY: Springer.

Mann, H. M. (2009). *Testing for differentially functioning indicators using mixtures of confirmatory factor analysis models.* (Unpublished doctoral dissertation). University of Maryland, College Park.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis, *Multivariate Behavioral Research, 33,* 181-220,

Masyn, K. E., & Henderson, C. E. (2010). Exploring the latent structures of psychological constructs in social development using the dimensional-categorical spectrum. *Social Development, 19, 3*.

McCutcheon, A. L. (1987). *Latent class analysis.* (Sage University Paper series on Quantitative Applications in the Social Sciences, No. 07-064). Newbury Park, CA: Sage.

McLachlan, G. J., Do, K.-A., & Ambroise, C. (2004). *Analyzing microarray gene expression data*. Hoboken, NJ: Wiley.

McLachlan, G., & Krishnan, T. (2008). *The EM algorithm and extensions.* New York, NY: John Wiley.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley and Sons.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39,* 479-515.

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

Muthén, B. (1998-2004). Mplus technical appendices. Los Angeles, CA: Muthén & Muthén.

Muthén, B. (2006). Should substance use disorders be considered as categorical or dimensional? *Addiction, 101*, 6-16.

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

Muthén, B. (2010). *Bayesian analysis in Mplus: A brief introduction* (Technical Report). Los Angeles, CA: Muthén & Muthén.

Muthén, B., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Mutiple-group and growth modeling in Mplus* (Mplus Web note No. 4). Retrieved June 30, 2010 from http://www.statmodel.com/mplus/examples/webnote.html

Muthén, B., & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors, 31*, 1050-1066.

Muthén B., & Asparouhov, T. (2009). Growth mixture modeling: Analysis with non-Gaussian random sets. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143-165). Boca Raton, FL: Chapman & Hall/CRC Press.

Muthén, B., & Asparouhov, T. (2010). *Bayesian SEM: A more flexible representation of substantive theory*. Manuscript submitted for publication.

Muthén, B., Asparouhov, T., Hunter, A., & Leuchter, A. (2010). *Growth modeling with non-ignorable dropout: Alternative analyses of the STAR*D antidepressant trial*. Manuscript submitted for publication.

Muthén, B., Asparouhov, T., & Rebollo, I. (2006). Advances in behavioral genetics modeling using Mplus: Applications of factor mixture modeling to twin data. *Twin Research and Human Genetics, 9,* 313-324.

Muthén, B., & Muthén, L. (1998-2010). *Mplus user's guide (6th ed.)* Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. (2010). Mplus version 6.1 [Computer software]. Los Angeles, CA: Muthén & Muthén.

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

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

Robert, C. P. (2007). *The Bayesian choice: From decision-theoretic foundations to computational implementation* (2nd ed). New York, NY: Springer Verlag.

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

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective* (Doctoral dissertation, University of Maryland). Available from ProQuest Dissertations and theses database. (UMI No. 3175148)

Segawa E., Emery S., & Curry S. (2008). Extended generalized linear latent and mixed model. *Journal of Educational and Behavioral Statistics, 33,* 464-484.

Skrondal, A., & Rabe-Hesketh, S. (2005). Structural equation modeling: Categorical variables. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1905-1910). New York, NY: Wiley.

Song, X.Y., Xia, Y. M., & Lee S. Y. (2009). Bayesian semiparametric analysis of structural equation models with mixed continuous and unordered categorical variables. *Statistics in Medicine, 28,* 2253-2276.

Spiegelhalter, D., Thomas, A., & Best, N. (2000). WinBUGS version 1.4 [computer program].

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Vermunt, J. K., & Magidson, J. (2005). Structural equation models: Mixture models. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1922-1927). Chichester, UK: John Wiley & Sons.

Von Davier, M. (2001). WINMIRA 2001: A software for estimating Rasch models, mixed and HYBRID Rasch models, and the latent class analysis [computer software]. Princeton, NJ: ETS.

Von Davier, M. (2005). mdltm [computer software]. Princeton, NJ: ETS.

Von Davier, M., & Rost, J. (2006) Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics,* vol. 26 (pp. 643-659), Amsterdam, The Netherland: Elsevier.

Yang, R., & Berger, J. O. (1994). Estimation of a covariance matrix using reference prior. *Annals of Statistics, 22,* 1195–1211.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods, 14,* 301-322.

Yung, Y. F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika, 62,* 297-330.