



THEMS  
2  
2006

LIBRARY  
Michigan State  
University

This is to certify that the  
dissertation entitled

ESTIMATING THE PARAMETERS FOR  
MULTIDIMENSIONAL ITEM RESPONSE THEORY MODELS  
BY MCMC METHODS

presented by

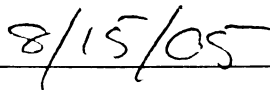
Yanlin Jiang

has been accepted towards fulfillment  
of the requirements for the

Ph. D degree in Education



Major Professor's Signature



Date

MSU is an Affirmative Action/Equal Opportunity Institution

**PLACE IN RETURN BOX** to remove this checkout from your record.  
**TO AVOID FINES** return on or before date due.  
**MAY BE RECALLED** with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE
AUG 09 2014		
0203 15		





**ESTIMATING PARAMETERS FOR MULTIDIMENSIONAL ITEM  
RESPONSE THEORY MODELS BY MCMC METHODS**

By

Yanlin Jiang

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Counselling, Educational Psychology and Special Education

2005

## ABSTRACT

# **ESTIMATING PARAMETERS FOR MULTIDIMENSIONAL ITEM RESPONSE THEORY MODELS BY MCMC METHODS**

By

Yanlin Jiang

Efforts to apply Markov Chain Monte Carlo (MCMC) methods to three-parameter linear logistic multidimensional IRT models are addressed using the Metropolis-Hastings algorithm within Gibbs approach. Bayesian modal estimators of both item and proficiency parameters are obtained in a simultaneous process rather than a separate parameter estimation procedure. It is shown that it is effective by blocking individual item discrimination and proficiency dimensional parameters and treating them without reference to other item and proficiency parameters. Both simple and complex structures of item dimensions are included. In addition, various proficiency dimensional structures are considered for three and five dimensional cases, respectively. The effects of four potential factors on model parameter estimation are investigated. Simulation studies are conducted across different designs for one-, three-, and five-dimensional cases. Results show that the parameter estimators based on MCMC are accurate in terms of correlation and root mean square errors. Numeric examples for the estimates of the standard errors demonstrate that the estimation is statistically stable and accurate.

## ACKNOWLEDGEMENTS

I am grateful for my dissertation committee: Dr. Mark Reckase (chair), Dr. Kimberly Maier, Dr. Richard Houang, and Dr. James Stapleton for their constructive comments and valuable suggestions. Without their inputs, this dissertation would not have been completed.

I would like to express my sincere gratitude to my academic advisor, Dr. Mark Reckase, for his constant support, direction, and encouragement over the past five years. I would also like to thank the Center for the Study of Curriculum and my supervisor, Dr. Richard Houang, whose final assistance supported the completion of the dissertation research and enabled the completion of my doctoral study. Working with him has been a tremendously rewarding experience for me.

Special thanks go to my husband, Deping Li, for his support, patience, and understanding in my life.

# Contents

<b>LIST OF TABLES . . . . .</b>	<b>vi</b>
<b>LIST OF FIGURES . . . . .</b>	<b>viii</b>
<b>1 Introduction . . . . .</b>	<b>1</b>
1.1 Item Response Theory Models . . . . .	1
1.1.1 The Uni-dimensional Item Response Theory Models . . . . .	1
1.1.2 The Multi-dimensional IRT Models . . . . .	4
1.2 Estimation Methods for IRT Models . . . . .	7
1.2.1 Commonly Used Estimation Methods and Their Limitations . . . . .	7
1.2.2 Applications of MCMC methods to Estimation of IRT-based Models . . . . .	12
1.3 The Importance of the Study . . . . .	13
<b>2 MCMC Methods for Parameter Estimation for Logistic MIRT Model . . . . .</b>	<b>17</b>
2.1 Overview of Markov Chain Monte Carlo Methods . . . . .	17
2.2 Likelihood Functions for the Linear Logistic MIRT Models . . . . .	21
2.3 M-H within Gibbs for Parameter Estimation for MIRT Models . . . . .	23
2.3.1 Complete Conditional Functions for Model Parameters . . . . .	23
2.3.2 Modelling the Covariance Structure for Multidimensional Abilities . . . . .	27
2.3.3 Random Walk Metropolis Algorithm within Gibbs . . . . .	31
2.4 Unbiased and Consistent Estimators of Parameters . . . . .	34
<b>3 Simulation Studies and Results . . . . .</b>	<b>36</b>
3.1 Prior Distributions for Model Parameters . . . . .	38
3.2 Diagnosing the Convergence of Markov Chains . . . . .	38
3.3 Initial Values and Iterations . . . . .	39

3.4	Estimating the Unidimensional 3PL Model . . . . .	41
3.4.1	Assessing Convergence . . . . .	42
3.5	Estimating the 3-Dimensional MIRT Model . . . . .	54
3.5.1	Generating Proficiency Parameters . . . . .	57
3.5.2	The Number of Proficiency Dimension and Sample Size . . . .	58
3.5.3	Proficiency Structure . . . . .	59
3.5.4	Generating Item Parameters . . . . .	60
3.5.5	The Estimation Accuracy and Stability for the 3-Dimensional MIRT Model . . . . .	62
3.6	Estimating the 5-dimensional Model . . . . .	69
3.7	Proficiency Structure Estimation . . . . .	82
3.8	Computing Time . . . . .	85
<b>4</b>	<b>Concluding Remarks and Future Research Directions</b>	<b>87</b>
	<b>BIBLIOGRAPHY</b>	<b>96</b>

# List of Tables

3.1	True Item Parameters for 30-Item Test (Dim = 1) . . . . .	43
3.2	True Item Parameters for 45-Item Test (Dim = 1) . . . . .	44
3.3	Estimates from three chains for 30-Item Test (Dim = 1, N = 2000) .	46
3.4	Item Parameter Estimates for 30-Item Test (Dim = 1) . . . . .	48
3.5	Item Parameter Estimates for 30-Item Test In BILOG-MG3 (Dim = 1)	49
3.6	Item Parameter Estimates for 45-Item Test (Dim = 1) . . . . .	50
3.7	Item Parameter Estimates for 45-Item Test (Dim = 1), cont. . . . .	51
3.8	RMSE for Estimating Uni-dimensional Models (Dim = 1) . . . . .	53
3.9	Correlations Between True Proficiency and Estimates (Dim = 1) . . .	54
3.10	True Item Parameters for 30-Item Test (Dim = 3) . . . . .	63
3.11	True Item Parameters for 45-Item Test (Dim = 3) . . . . .	64
3.12	RMSE for Multi-dimensional Test (Dim = 3, $\rho = .2$ ) . . . . .	64
3.13	RMSE for Multi-dimensional Test (Dim = 3, $\rho = \text{general}$ ) . . . . .	66
3.14	Correlations Between True Proficiency and Estimates (Dim = 3, $\rho = .2$ )	66
3.15	Correlations Between True Proficiency and Estimates (Dim = 3, $\rho =$ general) . . . . .	67
3.16	True Item Parameters for 30-Item Test (Dim = 5) . . . . .	75

3.17 True Item Parameters for 45-Item Test (Dim = 5) . . . . .	76
3.18 True Item Parameters for 45-Item Test (Dim = 5), cont. . . . .	77
3.19 RMSE for Multi-dimensional Test (Dim = 5, $\rho = .2$ ) . . . . .	80
3.20 RMSE for Multi-dimensional Test (Dim = 5, $\rho = \text{general}$ ) . . . . .	80
3.21 Correlations Between True Proficiency and Estimates (Dim = 5, $\rho = .2$ )	81
3.22 Correlations Between True Proficiency and Estimates (Dim = 5, $\rho =$ general) . . . . .	82
3.23 Estimates of Covariance Matrix, Dim = 3, $\rho = .2$ . . . . .	83
3.24 Estimates of Covariance Matrix, Dim = 3, $\rho = \text{general}$ . . . . .	83
3.25 Estimates of Covariance Matrix, Dim = 5, $\rho = \text{general}$ . . . . .	84
3.26 Estimates of Covariance Matrix, Dim = 5, $\rho = .2$ . . . . .	84
3.27 Computing time for 1-, 3-, and 5-Dimension data . . . . .	86
4.1 TESTFACT Item Parameters estimates for 30-Item Test (Dim = 3) .	97
4.2 TESTFACT Item Parameters Estimates for 30-Item Test (Dim = 5) .	98

# List of Figures

3.1	Sample ACF for series of $a_6$ , $Dim = 1$ . . . . .	45
3.2	Sample draw at first 3000 iterations for series of $a, b$ and $c$ . . . . .	47
3.3	True Proficiency Versus Estimates ( $Dim = 1$ ) . . . . .	55
3.4	True $a$ Parameter Versus Estimates ( $Dim = 1$ ) . . . . .	55
3.5	True $b$ Parameter Versus Estimates ( $Dim = 1$ ) . . . . .	56
3.6	True $c$ Parameter Versus Estimates ( $Dim = 1$ ) . . . . .	56
3.7	True Proficiency Versus Estimates ( $Dim = 3, \rho = general, n = 30, N = 5000$ ) . . . . .	69
3.8	True Proficiency Versus Estimates ( $Dim = 3, \rho = general, n = 45, N = 2000$ ) . . . . .	70
3.9	True Proficiency Versus Estimates ( $Dim = 3, \rho = general, n = 45, N = 2000$ ) . . . . .	70
3.10	True $a1$ Parameter Versus Estimates ( $Dim = 3, \rho = .2$ ) . . . . .	71
3.11	True $a2$ Parameter Versus Estimates ( $Dim = 3, \rho = .2$ ) . . . . .	71
3.12	True $a3$ Parameter Versus Estimates ( $Dim = 3, \rho = .2$ ) . . . . .	72
3.13	True $d$ Parameter Versus Estimates ( $Dim = 3, \rho = .2$ ) . . . . .	72



# **Chapter 1**

## **Introduction**

### **1.1 Item Response Theory Models**

Item response theory (IRT) becomes more and more important for psychological and educational testing. This philosophic and theoretic framework not only provides useful analytical tools (e.g., item differential functioning and test equating), but also provides an effective test design tool. The importance of the IRT framework cannot be realized unless the model parameters are accurately estimated given that the model assumptions are satisfied and the model is adequately fitted to the observed data.

In this chapter, both uni-dimensional and multi-dimensional logistic IRT models will be introduced, then some of the existing estimation methods will be reviewed, and finally the importance of a new method for estimating multidimensional IRT models will be addressed.

#### **1.1.1 The Uni-dimensional Item Response Theory Models**

Classical test theory (CTT) has been the mainstream of educational and psychological testing research and practice for many decades. Gulliksen's "Theory of Mental

Tests ” (1950) is one of the earliest books and a milestone of measurement theory. However, CTT suffers from a number of limitations, as is often seen in the literature (e.g., Embreston & Reise, 2000; Hambleton & Swaminathan, 1985). For example, item statistics (e.g., item difficulty) are sample dependent; reliability and standard errors of measurement estimators, which are the fundamental concepts in true score theory, do not take the proficiency differences among examinees into account. Hence, only a single reliability estimate is obtained for one test. Furthermore, CTT cannot probabilistically predict examinees’ response on items unless the items have previously been administered to similar individuals. In many testing contexts such as adaptive test, it is important to predict the examinee’s response in probability in order to provide next item for the examinee. As Lord states,

“we need to describe the items by item parameters and the examinees by examinee parameters in such a way that we can predict probabilistically the response of any examinees to any items, even if similar examinees have never taken similar items before (P.11, Lord, 1980)”.

Unfortunately, CTT fails to satisfy this property. Item response theory is a model-based measurement framework. IRT provides a more complete rationale for model-based measurement than CTT and overcomes a number of limitations of CTT (for details, please refer to Embreston & Reise, 2000). The important development of IRT is due to the work of Lord (1952, 1953), Birnbaum (1957, 1958a, 1958b), Lord and Novick (1968), and Rasch (1960). Various IRT-based models have been developed in the literature, for examples, the normal ogive models (Lord, 1952) and the logistic models (Rasch, 1960; Birnbaum, 1957, 1958a, 1958b, 1968; & Wright & Stone, 1979) for binary data, the graded response model (Samejima, 1969), the partial credit

model (Master, 1982), and the nominal response model (Bock, 1972) for polytomous data. There are other uni-dimensional IRT models (e.g., continuous response model, Samejima, 1972) but will not be discussed here since this study focuses on applying a new method to the logistic IRT models. One common feature of these models is that they explicitly predict the probability of correct response on an item given person and item parameters. More comparisons of other characteristics between CTT and IRT can be found in Embreston and Reise (2000).

In the family of IRT models, the three-parameter logistic model (3PL model) is one of the most widely used models. It was proposed by Birnbaum in 1968. For a dichotomous item, the item response function (IRF or called ICC) is the probability of a correct response to the item. This probability can be represented by the function (Lord, 1980)

$$p_i(\theta_j) \equiv p(U_{ij} = 1 \mid a_i, b_i, c_i, \theta_j) = c_i + (1 - c_i) \frac{\exp[1.7a_i(\theta_j - b_i)]}{1 + \exp[1.7a_i(\theta_j - b_i)]}, \quad (1.1)$$

where

$p_i(\theta_j)$  is the probability of correct answer to item  $i$  given the  $j$ th examinee's proficiency level  $\theta_j$ ;

$U_{ij}$  is the item response either 0 (incorrect) or 1 (correct) for examinee  $j$  on item  $i$ ;

$a_i$  is the  $i$ th item discriminating power; it is usually a positive number.

$b_i$  is the  $i$ th item difficulty;

$c_i$  is the  $i$ th item lower asymptote or called pseudo-guessing parameter; and

1.7 is a scale constant.

If there is no lower asymptote parameter in the above model, i.e.,  $c_i = 0$ , the 3PL model reduces to the 2PL model. Furthermore, if the discriminating power parameter  $a_i$  is treated as a constant in the model, then the model becomes 1PL model or Rasch model because of only one item parameter (i.e., item difficulty) in the model. Note that the 3PL, the 2PL, and the 1PL models only contain one proficiency parameter for each examinee, an important assumption for the models, which are labelled as uni-dimensional IRT models.

In addition to unidimensionality, another important assumption for IRT models is local independence. For a single examinee, the responses to the test items are related to each other only through this examinee's proficiency parameter(s). Hence, local independence can be understood as conditional independence. It assumes that examinee's responses to items are independent of each other after controlling for the examinee's proficiency parameter(s). The mathematical expression of local independence is given by

$$p(u_1, u_2, \dots, u_n \mid \theta) = \prod_{i=1}^n p_i(u_i \mid \theta), \quad (1.2)$$

where  $u_i$  is the item response on the  $i$ th item for a single examinee and  $i = 1, 2, \dots, n$ . Equation (1.2) implies that given a fixed proficiency parameter, the joint distribution  $p$  of responses to  $n$  items is the product of the marginal distributions  $p_i$  for all items.

### 1.1.2 The Multi-dimensional IRT Models

In the multi-dimensional item response theory (MIRT), items require multiple abilities to get a correct response. Under this circumstance, the uni-dimensional IRT models

are not adequate for such response data. A family of IRT models that contain multiple proficiency parameters is needed to reflect proficiency level on different dimensions for each examinee.

MIRT is an extension of uni-dimensional IRT. Like uni-dimensional IRT, MIRT models examinee's behavior (i.e., item response) given person and item characteristics. The essential difference of MIRT from uni-dimensional IRT is that in MIRT, multiple proficiency parameters are used to model person abilities and a vector form of item parameters to characterize items.

To describe MIRT-based models, it is necessary to introduce the concept *complete latent space*. Lord defined it as a collection of all those latent variables  $\theta_k$ 's that discriminate among groups of examinees (Lord & Novick, 1968) for  $k = 1, 2, \dots, p$ , where  $p$  is the number of proficiency dimensions. Denote the *complete latent space*  $\boldsymbol{\theta}$  by the vector

$$\boldsymbol{\theta} \equiv (\theta_1, \theta_2, \dots, \theta_p)'. \quad (1.3)$$

These variables can be thought of as “psychological dimensions necessary for the psychological description of individuals” (p.359). For the population of examinees, every single examinee possesses a value for each of the latent variables in the space. For uni-dimensional IRT models, the *complete latent space* has only one variable. For multi-dimensional IRT models, it is assumed that two or more latent variables are needed to characterize an examinee's proficiency.

There are a few MIRT-based models. Early MIRT models for binary data were from the work of McDonald (1967) and Lord & Novick (1968). Other models have

also been found in the literature. For example, the multidimensional Rasch model (Stegmann, 1983), the multidimensional two-parameter normal ogive IRT model (Bock, Gibbons and Muraki, 1988), the multicomponent latent trait model (MLTM; Whitely, 1980), etc. Reckase provides the extension of the uni-dimensional three-parameter logistic model to multi-dimensional form (Reckase, 1985, 1996). He pointed out that

“After reviewing many possible models that include vector parameters for both examinee and item characteristics [see McKinley and Recakse (1982) for a summary], the model given below was selected for further development because it was reasonable given what is known about item response data, consistent with simpler, uni-dimensional item response theory models, and estimable with commonly attainable numbers of examinees and test items (p.272)”.

$$p_i(\boldsymbol{\theta}_j) \equiv p(U_{ij} = 1 \mid \mathbf{a}_i, d_i, c_i, \boldsymbol{\theta}_j) = c_i + (1 - c_i) \frac{\exp(\mathbf{a}_i' \boldsymbol{\theta}_j + d_i)}{1 + \exp(\mathbf{a}_i' \boldsymbol{\theta}_j + d_i)}, \quad (1.4)$$

where

$p(U_{ij} = 1 \mid \mathbf{a}_i, d_i, c_i, \boldsymbol{\theta}_j)$  is the probability of a correct response (score of 1) for examinee  $j$  on test item  $i$ ;

$U_{ij}$  is a dichotomous random variable representing the item response for examinee  $j$  on item  $i$ ;

$\boldsymbol{\theta}_j$  is the vector of abilities for examinee  $j$ , i.e.,  $\boldsymbol{\theta}_j \equiv (\theta_{j1}, \theta_{j2}, \dots, \theta_{jp})'$ ;

$\mathbf{a}_i$  is a vector of parameters related to the discriminating power of the test item  $i$  (the rate of change of the probability of correct response to changes in trait levels for the examinees);

$d_i$  is a parameter related to the difficulty of item  $i$ ;

$c_i$  is the probability of correct response that is approached when the abilities assessed by item  $i$  are very low;  $c_i$  is usually called the lower asymptote, or less correctly, the guessing parameter.

The unique contribution of the model above, as summarized by Recakse (1997), is that it focuses on the characteristics of the test items and the way they interact with the examinee population. This model has proved to be useful for a variety of applications and has helped in conceptualizing a number of psychometric problems including the assessment of differential functioning and test parallelism (Ackerman, 1990, 1992).

## **1.2 Estimation Methods for IRT Models**

### **1.2.1 Commonly Used Estimation Methods and Their Limitations**

IRT models contain at least two types of parameters: person parameters (also called latent trait, proficiency, or ability parameters) and item parameters. Estimating person parameters for IRT models is frequently accomplished by using one of three methods: (1) maximum likelihood (ML); (2) maximum a posteriori (MAP); and (3) expected a posteriori (EAP). The ML method estimates person parameters by maximizing the likelihood of an examinee's item responses. But one critical problem in the ML method is that the ML cannot estimate person parameters for examinees who have all correct or all incorrect response patterns (p.162, Embreston & o Reise, 2000). In addition, ML estimates have the consistency property only as sample size

increases (here sample size refers to the number of test items, or test length), which in reality, is not an easy condition to meet because the test is often viewed as a fixed set of items.

Both EAP and MAP are from the Bayesian perspective. MAP (also called Bayesian Modal Estimation) scoring method uses prior information about person proficiency in conjunction with the likelihood function to estimate proficiency level by maximizing a posterior distribution. The advantage of MAP is that proficiency can be estimated for all possible response patterns including perfect pattern. The perfect pattern could be all-correct response pattern, all-incorrect response pattern, or some odd pattern that makes it difficult for the ML procedure to find solutions (e.g., no solution, or multiple solutions). Critics of Bayesian modal estimation methods is the proficiency estimates may depend on heavily the choice of the prior distribution of proficiency parameters especially when the sample size (i.e., test length) is small. EAP is a method of finding the mean of a posterior distribution. One advantage of the EAP estimator is that it “has minimum mean square error over the population of ability” (p.439, Bock & Mislevy, 1982). However, the estimates from EAP are biased (Wainer & Thissen, 1987).

Item parameters in IRT models are usually estimated by the maximum likelihood (ML) approach. The commonly used methods under this approach are (a) joint maximum likelihood (JML), (b) marginal maximum likelihood (MML), and (c) conditional maximum likelihood (CML).

It is known that the consistency property of the maximum likelihood estimator



holds for person parameters only when item parameters are known and the number of items increases. Similarly, the consistent item parameter estimates can be obtained when person parameters are known and the number of examinees increases. The JML procedure simultaneously estimates person and item parameters for all items and examinees by jointly maximizing the likelihood function of the response data. In principle, this procedure is straightforward. However, it has several drawbacks in practice as some researchers pointed out. First, nonlinear (i.e., S-shape) item characteristic curve (ICC) results in nonlinear likelihood equations. Solving nonlinear equation systems is often a formidable task (Hambleton & Swaminathan, 1985). Secondly, when used with the 3PL model, large numbers of examinees (e.g., more than 1000) are required for accurate item parameter estimation (e.g., Lord & Novick, 1968; Swaminathan & Gifford, 1979). Thirdly, increasing the number of examinees cannot guarantee the estimation improvement (Hulin, Lissak, & Drasgow, 1982). That is, the consistency property of estimation does not always hold due to increase in both item (structure) and person (incidental) parameters simultaneously.

When sufficient statistics are available for person parameters, one may avoid the problem of presenting person parameters in the likelihood function. For the Rasch model, since the number correct score (also called total score) is a sufficient statistic for the proficiency parameter, it is possible to express the likelihood function  $L(U \mid \theta, b_i)$  in terms of total score instead of proficiency parameters. The CML procedure can be used to estimate item parameters and the corresponding estimates are consistent (Hambleton & Swaminathan, 1985). However, since CML requires a sufficient statistic

for estimating trait level, it is restricted to the Rasch model family. In more complex models such as the 2PL, the 3PL and the MIRT models, proficiency estimates are dependent on item characteristics. Therefore the total score is no longer a sufficient statistic for estimating proficiency. In addition, Embreston and Reise (2000) pointed out several other disadvantages on CML estimation procedure: no estimates for items or persons are available for perfect response pattern (P.218); numerical problems often occur for long tests, complicated patterns of missing data, or polytomous data.

Estimating item parameters can be carried out if the likelihood function can be expressed without any reference to the person parameters. Assuming the underlying distribution of proficiency is continuous and known, the essence of MML is to integrate over the proficiency distribution, then the item parameters are estimated in the marginal distribution (Bock & Lieberman, 1970). This procedure removes the dependency of item parameter estimates on the proficiency estimates. The advantage of MML is its estimates possess the consistency property since increasing number of examinees doesn't require additional estimation of proficiency estimates (Kiefer & Wolfowitz, 1956). The MML approach is accomplished within the framework of the EM algorithm (p.190, Baker, 1992). Although MML/EM has lot of nice features and becomes a standard for item parameter estimation, Baker (p.190, Baker, 1992) pointed out that certain limitations of this approach exist in practice. For example, items that are answered correctly or incorrectly by all examinees have to be eliminated for item parameter estimation before calibration, an obvious loss of data information; certain data set can yield large absolute value of item difficulty and other deviant

values as item parameter estimates. Once these deviant values are used for proficiency estimation, it will cause estimation process to fail. In addition, although many has done research on an accelerated EM algorithm which is faster, the EM algorithm convergence rate is slow when estimating high-dimensional models.

If prior information about item parameters is available, Bayesian estimation methods are possible for IRT-based models. In 1982, 1985, and 1986, Swaminathan and Gifford (1982, 1985, 1986) derived Bayesian estimation procedures for the one-, two-, and three-parameter logistic models, where item parameter estimation takes place without any marginalization. Mislevy (1986b), Tsutakawa and Lin (1986) took a different approach, which inherited properties of MML by integrating (i.e., marginalizing) proficiency parameter out of likelihood function. Marginal Bayesian modal estimation is accomplished within the framework of the EM algorithm (Baker, 1992) too. However, marginalized Bayesian item parameter estimates may heavily depend on the item priors in particular for small sample size, and hence the resulting item parameter estimates will be shrunk to the mode of its corresponding prior distribution for informative priors.

The frequently used estimation methods and their limitations are summarized in this section. For one-dimensional IRT models, although joint maximum likelihood estimates are available in some programs to estimate item and proficiency parameters simultaneously (e.g., LOGIST uses joint maximum likelihood estimation paradigm formulated by Alan Birnbaum in 1968), the estimates of proficiency parameters need not be consistent as the sample size increases (e.g., Neyman & Scott, 1948; Little &

Rubin, 1983). In addition, in some extreme situations of responses, the maximum likelihood procedure could give positive or negative infinity estimates for proficiency parameters.

MML/EM procedure has become a central methodology for parameter estimates in the IRT framework. However, when test settings get more complex (e.g., with presence of missing data and polytomously score data) and IRT models are more complicated (e.g., the MIRT models), application of EM algorithm becomes less straightforward (Patz & Junker, 1999a).

In Section 1.3, the importance of a new method for parameter estimation in linear logistic MIRT models will be addressed.

### **1.2.2 Applications of MCMC methods to Estimation of IRT-based Models**

A new estimation approach that could avoid some shortcomings of the estimation procedures discussed above is desired to improve the estimation accuracy in particular for the more complicated testing practices and the complex IRT models. Markov Chain Monte Carlo (MCMC) methods, which are from a Bayesian perspective, can be applied to estimating parameters for IRT models.

Researchers have had interests in MCMC methods for several decades (e.g., Metropolis, et al., 1953). MCMC methods have been successful in many Bayesian applications because they allow one to draw samples from a wide range of interested posterior distributions, including many for which simulation methods were previously much more difficult to implement ( e.g., Gilks, Richardson, & Spiegelhalter, 1996).

MCMC methods have also been recently implemented for parameter estimation and inference through stochastic simulation for IRT models. Patz and Junker (1999a) demonstrate that MCMC techniques are well-suited to complex models with IRT assumptions and the MCMC methodology can be routinely implemented to fit the IRT contexts, and further address the strategies and issues of extending the basic MCMC methods for Bayesian inference in complex IRT settings such as non-response, designed missingness, multiple raters, guessing behaviors, and partial credit (i.e., polytomous) test items (Patz & Junker, 1999b). Earlier work can trace back to Albert (1992), who estimated the two-parameter normal ogive model for augmented data using the Gibbs sampler. Various applications of MCMC methods have also been developed in the literature for item parameters recovery (e.g., Wollack, Bolt, Cohen, & Lee, 2002; Mathews & Hombo, 2001; Kim & Cohen, 1998; De-la-Torre, Patz, 2001; Maris & Maris, 2002; Fox, 2002; Williamson, Johnson, Sinharay & Bejar, 2002), for coefficient alpha estimates (Li & Woodruff, 2001), etc.

Different from the Bayesian modal estimates discussed in Section 1.2, the MCMC estimates of parameters will no longer be dependent on the prior distribution and the parameter estimates are not shrunk to the mean of prior distribution.

### **1.3 The Importance of the Study**

Recently, Segall (1996, 2001) has advanced multidimensional adaptive testing (MAT) and the measure of general proficiency using a linear logistic MIRT model. He found

that MAT could provide equal or higher reliability with fewer items than are required in one-dimensional adaptive test. He concludes that in addition to increasing measurement efficiency, MAT can also be used as a tool ensuring adequate and efficient coverage of content for examinees at different levels of proficiency (Segall, 1996). However, as he emphasizes, further study is needed before MAT can be routinely applied and item parameter estimation for MIRT models must be refined.

In estimating parameters for MIRT models, simple structure (i.e., each item only measure one dimension of proficiency) is sometimes assumed (e.g., De-la-Torre, Patz, 2001). the Multi-unidimensional approach, as suggested by Segall (e.g., 1996), is an example of a simple structure. In this approach, several sub-tests measuring different contents are given at one test administration. There are two ways to estimate the model parameters for the multi-unidimensional approach. One is estimating the model parameter for the tests separately (i.e., independently), which is not realistic since usually the contents to be measured are correlated. The other way is to treat each content as one dimensional, then estimate the model parameters simultaneously using a multidimensional model. Segall (1996) pointed out that although the multi-unidimensional approach is appealing in terms of its simple structure, it may suffer at least two undesirable features. One may be due to the poor specification of the elements of the covariance matrix of the proficiency vector, and the other is that the assumption of simple structure may lead to some poorly specified loadings (p.350). In addition, to develop a common metric and orientation of item parameter estimates for MIRT models is not convenient or even unlikely to be achieved. Segall

(1996) addresses that when developing large item pools with several dimensions, it is often necessary to divide the pools into subsets of items. This design however may raise several issues concerning the metric of the latent dimensions. Therefore, a new methodology is desirable for the concurrent estimation of item parameters for MIRT models for building item pool before MAT can be more reliably implemented.

Both item and proficiency parameters in MIRT models can be estimated simultaneously using MCMC methods. Parameter estimation using MCMC methods is different from a number of approaches for estimating MIRT models (Carlson, 1987; Fraser, 1988; McDonald, 1985; McKinley & Reckase, 1983; Muthen, 1984). Efforts to apply MCMC methods to multidimensional models have been explored in the literature. For example, Beguin and Glas (1998) generalized the Albert (1992) procedure to the unidimensional 3PL normal ogive model and  $Q$ -multidimensional normal ogive models. However, the study assumes the underlying covariance matrix for abilities is an identity matrix, which is not realistic since the proficiency dimensions in one test are more likely to be correlated. Moreover, the values of item parameters in the study are restricted to a small range (e.g.,  $a$  is from 0 to 1,  $d$  is from -1 to 1), which is also not realistic for a general and more complex testing context.

De-la-Torre and Patz (2001) examine simultaneous proficiency estimation for MIRT models using MCMC approach. But the study only assumes the simple structure. In addition, to estimate the proficiency parameters, the study assumes the item parameters are known, which actually is not available in many applications.

Bolt and Lall (2003) investigate the item parameter estimation of compensatory

and noncompensatory MIRT models using the MCMC method. In their study, the guessing parameter was not included in the MIRT models and only two-dimensional model was considered. In addition, the item parameters cover only a small range of values.

However, not much attention has been paid to three-parameter MIRT models that has been proven useful for a variety of applications in the literature. It is necessary to study parameter estimation using MCMC methods in a more general, complex, and realistic situations. For example, guessing parameter is included to the model, complex item dimension structures (i.e., each item measures one dimension or more than one dimension of abilities) are considered in the test design with an exploratory solution, and the inter-correlation among proficiency dimensions will be estimated and not limited to the identity matrix or special pattern of covariance matrix (e.g., all off-diagonal elements are the same). Moreover, the current study intends to examine the impact of four factors – the test length, the number of dimensions, the sample size, and the proficiency covariance structure on the accuracy and stability of parameter estimates for MIRT models.



## Chapter 2

# MCMC Methods for Parameter Estimation for Logistic MIRT Model

### 2.1 Overview of Markov Chain Monte Carlo Methods

Statistical inference is a procedure for drawing conclusions about population parameters from the observed sample data. Bayesian statistical conclusions about a parameter are typically made in terms of a probability statement conditioned on the observed data, or the posterior of the interested parameter. A sample generated by MCMC methods can be used for statistical inference, including point estimate, the construction of a marginal density, prediction, estimation of moments, and so on.

Gill (2002) defined *Markov chain* as:

“a stochastic process with the property that any specified state in the series,  $\theta^{(t)}$ , is only dependent on the previous value of the chain. Or in a probability expression (p.302):

$$P(\theta^{(t)} \in A \mid \theta^{(0)}, \theta^{(1)}, \dots, \theta^{(t-2)}, \theta^{(t-1)}) = P(\theta^{(t)} \in A \mid \theta^{(t-1)}), \quad (2.1)$$

Where  $A$  is an event or range of events in the complete state space;  $t$  is a positive

number referring to the  $t$ th time interval;  $\theta$  is a random quantity taking values in some known state space,  $\theta$ .

The Monte Carlo method uses random samples from the desired distribution instead of calculating quantities from the analytical form to summarize the interested theoretical distribution.

Generally speaking, the Markov Chain Monte Carlo methods involve two steps. First, producing a chain in which each value only depends on the previous value. Second, once this chain converges to the desired posterior distribution, the Monte Carlo method is used to summarize the interested distribution.

There are two basic methods in MCMC: (1) Gibbs sampler; (2) Metropolis-Hastings algorithm.

The Gibbs sampler named by Geman and Geman (1984) is one of the most widely used MCMC techniques. Let  $Q$  be the model parameters vector with  $k$  components, and  $q_i$  be the  $i$ th model parameter in  $Q$ . Denote  $Q \equiv (q_1, q_2, \dots, q_i, \dots, q_k)$  and  $Q_{-i} \equiv (q_1, q_2, \dots, q_{i-1}, q_{i+1}, \dots, q_k)$ . Then  $Q$  can be expressed as  $Q \equiv Q_{-i} \cup q_i$ . Denote the complete conditional function of the  $i$ th parameter by  $P(q_i | Q_{-i}) \equiv P(q_i | q_1, q_2, \dots, q_{i-1}, q_{i+1}, \dots, q_k)$ .

The *Gibbs sampler* sequentially samples from the complete conditional distributions  $P(q_i | Q_{-i}, y), i = 1, \dots, k$ , where  $y$  indicates observed data.

Then Gibbs sampling algorithm can be defined as the following:

1. Specify the starting values for the model parameter vector  $Q$ , i.e.,

$$Q^{(t=0)} = (q_1^{(t=0)}, q_2^{(t=0)}, \dots, q_k^{(t=0)}).$$

2. At  $t + 1$ th iteration, simulate

$$q_1^{(t+1)} \text{ from } p(q_1 \mid q_2^{(t)}, q_3^{(t)}, \dots, q_k^{(t)})$$

$$q_2^{(t+1)} \text{ from } p(q_2 \mid q_1^{(t+1)}, q_3^{(t)}, \dots, q_k^{(t)})$$

$\vdots$

$$q_i^{(t+1)} \text{ from } p(q_i \mid q_1^{(t+1)}, q_2^{(t+1)}, \dots, q_{i-1}^{(t+1)}, q_{i+1}^{(t)}, \dots, q_k^{(t)})$$

$\vdots$

$$q_k^{(t+1)} \text{ from } p(q_k \mid q_1^{(t+1)}, q_2^{(t+1)}, \dots, q_{k-1}^{(t+1)}) \text{ sequentially.}$$

3. Set  $t = t + 1$  and repeat step 2 until convergence.

The second frequently used method is the *Metropolis-Hastings algorithm* (M-H algorithm, Metropolis et al, 1953; Hastings, 1970). This method is applied when it is difficult to simulate from the complete conditional distributions by traditional methods (by the method of rejection sampling or by a known generator, for example).

A Markov chain using the M-H algorithm can be obtained as follows:

For any parameter  $\theta$ ,

1. Assign an initial value for parameter  $\theta$ .
2. Specify a proposal density  $r(\theta^t, \theta^{(t+1)})$ , which defines the proposal density from state  $\theta^t$  to state  $\theta^{(t+1)}$ .

3. Given the current state  $\theta^t$ , the candidate  $\theta^*$  for the next state  $\theta^{(t+1)}$  in the chain is sampled from  $r(\theta^t, \theta^{(t+1)})$ .
4.  $\theta^*$  is accepted as the next value  $\theta^{(t+1)}$ , i.e.,  $\theta^{(t+1)} = \theta^*$  with probability  $\alpha(\theta^t, \theta^*)$ , where

$$\alpha(\theta^t, \theta^*) = \min \left\{ \frac{g(\theta^*)r(\theta^*, \theta^t)}{g(\theta^t)r(\theta^t, \theta^*)}, 1 \right\}, \quad (2.2)$$

and  $g(\cdot)$  is the density of the target distribution.

5. If  $\theta^*$  is rejected, then the next value will stay at current state, i.e., assign  $\theta^{(t+1)} = \theta^t$ .

The M-H algorithm first simulates a Markov chain whose distribution differs from the desired distribution for the parameter, and then subsequently uses the acceptance probability to reject or accept the value such that a new Markov chain is constructed that has the target posterior as its stationary distribution.

It has been shown that the Gibbs sampler is a special case of the M-H algorithm where the probability of accepting the candidate value is always one (p.436, Gelman 1992; p.182, Tanner 1996). The distinction between the Gibbs sampler and the M-H algorithm is that the M-H algorithm requires the complete conditional distribution and so it is more restrictive (p.166, Gamerman 1997, Besag et al. 1995, Tierney 1991).

The combination of the Gibbs sampler and the M-H algorithm is a hybrid algorithm. One value is generated from the M-H procedure, followed by the next Gibbs step. Like the Gibbs sampler and the M-H algorithm, the M-H within the Gibbs

algorithm also produces a Markov chain with the correct stationary distribution.

## 2.2 Likelihood Functions for the Linear Logistic MIRT Models

If pre-calibrated item parameters are available, maximum likelihood estimates or Bayesian modal estimates of the proficiency parameters can be obtained. Suppose the assumption of local independence is held for the MIRT models. Then the probability of a set of observed responses  $\mathbf{u}_j = (u_{1j}, u_{2j}, \dots, u_{ij}, \dots, u_{nj})$  for the  $j$ th examinee with proficiency vector  $\boldsymbol{\theta}_j$  on  $n$  items is equal to the product of the probabilities associated with the response to each item.

$$L(\mathbf{u}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{d}, \mathbf{c}) = p(u_{1j}, u_{2j}, \dots, u_{ij}, \dots, u_{nj} \mid \boldsymbol{\theta}_j) \quad (2.3)$$

$$= \prod_{i=1}^n p_i(\boldsymbol{\theta}_j)^{u_{ij}} (1 - p_i(\boldsymbol{\theta}_j))^{1 - u_{ij}}, \quad (2.4)$$

where

$u_{ij}$  is a response (0 or 1) of the  $j$ th examinee on the  $i$ th item;

$\boldsymbol{\theta}_j$  is a  $p$ -dimensional proficiency vector, i.e.,  $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \dots, \theta_{jp})$ .

$p_i(\boldsymbol{\theta}_j)$  is the probability of the  $j$ th examinee correctly answering the  $i$ th item. Similarly, the probability of a set of  $N$  observed responses

$\mathbf{v}_i = (v_{i1}, v_{i2}, \dots, v_{ij}, \dots, v_{iN})$  for the  $i$ th item is given by

$$\begin{aligned} L(\mathbf{v}_i \mid \boldsymbol{\Theta}, \boldsymbol{\Sigma}, \mathbf{a}_i, \mathbf{d}_i, \mathbf{c}_i) &= \prod_{j=1}^N p(v_{i1}, v_{i2}, \dots, v_{ij}, \dots, v_{iN} \mid \boldsymbol{\Theta}, \boldsymbol{\Sigma}, \mathbf{a}_i, \mathbf{d}_i, \mathbf{c}_i) \\ &= \prod_{j=1}^N p_i(\boldsymbol{\theta}_j)^{u_{ij}} (1 - p_i(\boldsymbol{\theta}_j))^{1 - u_{ij}}, \end{aligned}$$

According to Bayes theorem, the posterior density function of  $\theta_j$  for  $j = 1, 2, \dots, N$ , can be expressed as

$$f(\theta_j | \mathbf{u}_j) = L(\mathbf{u}_j | \theta_j) \frac{\pi_{\theta}(\theta_j)}{m(\mathbf{u}_j)} \propto L(\mathbf{u}_j | \theta_j) \pi_{\theta}(\theta_j) \quad (2.5)$$

where

$L(\mathbf{u}_j | \theta_j)$  is the likelihood function given by (2.3);

$\pi_{\theta}$  is the prior distribution of  $\theta$ ;

$m(\mathbf{u}_j)$  is the marginal probability density of  $\mathbf{u}_j$ ; and

$N$  is the number of examinees.

Assume the prior distribution of  $\theta$  is a multivariate normal with mean vector  $\mu$  and the covariance matrix  $\Sigma$ , then the density of  $\pi_{\theta}(\theta_j)$  is

$$\pi_{\theta}(\theta_j) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\theta_j - \mu)' \Sigma^{-1}(\theta_j - \mu)\right]. \quad (2.6)$$

Maximizing  $L(\theta_j | \mathbf{u}_j)$  can obtain the Bayesian modal estimates of an individual proficiency parameter vector  $\theta_j, \forall j = 1, 2, \dots, N$ . That is to solve the equations as

$$\frac{\partial}{\partial \theta_{jk}} \log L(\theta_j | \mathbf{u}_j) = 0, \forall k = 1, 2, \dots, p; j = 1, 2, \dots, N. \quad (2.7)$$

Nevertheless, in many applications, the item parameters are not available, or both item and proficiency parameters are required to estimate from the observed data. The following section is to address the simultaneous estimation of the item and proficiency parameters using the MCMC methods.

## 2.3 M-H within Gibbs for Parameter Estimation for MIRT Models

### 2.3.1 Complete Conditional Functions for Model Parameters

Under the assumption of local independence, the overall likelihood function of responses for  $N$  examinees on  $n$  items can be written as

$$\begin{aligned}
 L(\mathbf{U} \mid \boldsymbol{\Theta}, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{d}, \mathbf{c}) &= \prod_{j=1}^N L(\mathbf{u}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{d}, \mathbf{c}) \\
 &= \prod_{i=1}^n L(\mathbf{v}_i \mid \boldsymbol{\Theta}, \boldsymbol{\Sigma}, \mathbf{a}_i, d_i, c_i) \\
 &= \prod_{j=1}^N \prod_{i=1}^n p_i(\boldsymbol{\theta}_j)^{u_{ij}} (1 - p_i(\boldsymbol{\theta}_j))^{1-u_{ij}},
 \end{aligned}$$

where

$\boldsymbol{\Theta}$  is a  $N \times p$  matrix representing all proficiency parameters, i.e.,

$$\boldsymbol{\Theta} \equiv (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_j, \dots, \boldsymbol{\theta}_N)' = \begin{pmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1p} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{N1} & \theta_{N2} & \cdots & \theta_{Np} \end{pmatrix};$$

$\boldsymbol{\Sigma}$  is a  $p \times p$  variance-covariance matrix for  $\boldsymbol{\theta}_j$  under the assumption that each examinee comes from a multivariate normal population, i.e.,  $\boldsymbol{\theta} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ ;  $p$  is the number of dimensions of proficiency parameters;

$\mathbf{A}$  is a  $n \times p$  matrix representing all  $a$  parameters for  $n$  items, i.e.,

$$\mathbf{A} \equiv (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_i, \dots, \mathbf{a}_n)' = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix};$$

$\mathbf{a}_i$  is a row vector with  $p$  components representing all of parameters related to discriminating power for the  $i$ th item, i.e.,  $\mathbf{a}_i = (a_{i1}, a_{i2}, \dots, a_{ip})$ ;

$\mathbf{d}$  is a vector of  $d$  parameters for  $n$  items;

i.e.,  $\mathbf{d} \equiv (d_1, d_2, \dots, d_i, \dots, d_n)'$ ;

$\mathbf{c}$  represents a vector of all pseudo-guessing parameters for a test of  $n$  items,

i.e.,  $\mathbf{c} \equiv (c_1, c_2, \dots, c_i, \dots, c_n)'$ ;

$\mathbf{U}$  is a  $N \times n$  matrix of responses data for all  $N$  examinees on  $n$  items;

$\mathbf{v}_i$  is a response vector for all  $N$  examinees on the  $i$ th item,

i.e.,  $\mathbf{v}_i \equiv (v_{i1}, v_{i2}, \dots, v_{ij}, \dots, v_{iN})'$ .

$\mathbf{u}_j$  is a row vector for the  $j$ th examinee's response on all items,

i.e.,  $\mathbf{u}_j \equiv (u_{1j}, u_{2j}, \dots, u_{nj})$ .

The above equations tell that the likelihood function for the  $N \times n$  response matrix can be expressed as either the product of the likelihood functions across all examinees or the product of the likelihood functions across all items.

Let  $\pi(\boldsymbol{\Theta}, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{d}, \mathbf{c})$  denote the joint prior distribution of all parameters in the



model  $\Theta, \Sigma, \mathbf{A}, \mathbf{d}$ , and  $\mathbf{c}$ . Assume that the prior distributions of both item and person parameters are independent. Then the joint prior distribution can be written as

$$\begin{aligned}\pi(\Theta, \Sigma, \mathbf{A}, \mathbf{d}, \mathbf{c}) &= \pi_{\Theta}(\Theta \mid \Sigma)\pi_{\Sigma}(\Sigma)\pi_{\mathbf{A}}(\mathbf{A})\pi_{\mathbf{d}}(\mathbf{d})\pi_{\mathbf{c}}(\mathbf{c}) \\ &= \left(\prod_{j=1}^N \pi_{\theta}(\theta_j \mid \Sigma)\pi_{\Sigma}(\Sigma)\right) \prod_{i=1}^n \pi_{\mathbf{a}}(\mathbf{a}_i)\pi_d(d_i)\pi_c(c_i).\end{aligned}$$

The joint posterior function for model parameters can be expressed as

$$p(\Theta, \Sigma, \mathbf{A}, \mathbf{d}, \mathbf{c} \mid \mathbf{U}) \propto L(\mathbf{U} \mid \Theta, \Sigma, \mathbf{A}, \mathbf{d}, \mathbf{c})\pi(\Theta, \Sigma, \mathbf{A}, \mathbf{d}, \mathbf{c})$$

Apparently, we cannot simulate samples from the joint posterior distribution directly, since the joint posterior is not a known distribution for direct sampling.

In order to sample values for the model parameters from the joint posterior distribution, the Metropolis-Hastings within Gibbs (Gibbs/M-H) algorithm is implemented, which is found to be effective in experimenting with new models (Patz & Junker, 1997), the complete conditional distributions of the parameters in MIRT models are analytically expressed in the following:

The complete conditional distribution of the proficiency parameters by Bayes the-

orem is

$$\begin{aligned}
P_{\theta}(\theta_j \mid \Theta_{-j}, \Sigma, \mathbf{A}, \mathbf{d}, \mathbf{c}, \mathbf{U}) &= P(\theta_j \mid \mathbf{u}_j, \Sigma, \mathbf{A}, \mathbf{d}, \mathbf{c}) \\
&\propto L(\mathbf{u}_j \mid \theta_j, \mathbf{A}, \mathbf{d}, \mathbf{c}) \pi_{\theta}(\theta_j) \\
&= \prod_{i=1}^n p_i(\theta_j)^{u_{ij}} (1 - p_i(\theta_j))^{1-u_{ij}} \pi_{\theta}(\theta_j),
\end{aligned}$$

where  $\Theta_{-j} = (\theta_1, \theta_2, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_N)'$ . Note that  $\Theta = \theta_j \cup \Theta_{-j}$ .

Similarly, we can have the complete conditional distributions for item parameters.

That is,

$$\begin{aligned}
P_{\mathbf{a}}(\mathbf{a}_i \mid \mathbf{A}_{-i}, \Theta, \Sigma, \mathbf{d}, \mathbf{c}, \mathbf{U}) &= \frac{P(\mathbf{v}_i \mid \mathbf{a}_i, \Theta, \Sigma, d_i, c_i) P(\mathbf{a}_i, \Theta, \Sigma, d_i, c_i)}{P(\Theta, \Sigma, \mathbf{v}_i, d_i, c_i)} \\
&\propto L(\mathbf{v}_i \mid \Theta, \mathbf{a}_i, d_i, c_i) \pi_{\mathbf{a}}(\mathbf{a}_i) \\
&= \prod_{i=1}^n p_i(\theta_j)^{u_{ij}} (1 - p_i(\theta_j))^{1-u_{ij}} \pi_{\mathbf{a}}(\mathbf{a}_i),
\end{aligned}$$

It can be shown that the complete conditional distributions for  $d_i$  and  $c_i$  have the following expressions:

$$\begin{aligned}
P_d(d_i \mid \mathbf{d}_{-i}, \Theta, \Sigma, \mathbf{A}, \mathbf{c}, \mathbf{U}) &\propto L(\mathbf{v}_i \mid \Theta, \mathbf{a}_i, d_i, c_i) \pi_d(d_i) \\
&= \prod_{i=1}^n p_i(\theta_j)^{u_{ij}} (1 - p_i(\theta_j))^{1-u_{ij}} \pi_d(d_i),
\end{aligned}$$

$$\begin{aligned}
P_c(c_i \mid \mathbf{c}_{-i}, \Theta, \Sigma, \mathbf{A}, \mathbf{d}, \mathbf{U}) &\propto L(\mathbf{v}_i \mid \Theta, \mathbf{a}_i, d_i, c_i) \pi_c(c_i) \\
&= \prod_{i=1}^n p_i(\theta_j)^{u_{ij}} (1 - p_i(\theta_j))^{1-u_{ij}} \pi_c(c_i),
\end{aligned}$$

where

$\pi_{\mathbf{a}}$ ,  $\pi_d$ , and  $\pi_c$  are the prior distributions for  $\mathbf{a}$ ,  $d$ , and  $c$  respectively;

$\mathbf{A}_{-i}$  is a  $(n - 1) \times p$  matrix, i.e.,  $\mathbf{A}_{-i} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{i-1}, \mathbf{a}_{i+1}, \dots, \mathbf{a}_n)$ ;

$\mathbf{d}_{-i}$  is a vector with  $(n - 1)$  components, i.e.,  $\mathbf{d}_{-i} = (d_1, d_2, \dots, d_{i-1}, d_{i+1}, \dots, d_n)$ ;

$\mathbf{c}_{-i}$  is a vector with  $(n - 1)$  components, i.e.,  $\mathbf{c}_{-i} = (c_1, c_2, \dots, c_{i-1}, c_{i+1}, \dots, c_n)$ ; and

$p_i(\theta_j)$  is as previously defined in equation (1.4).

### 2.3.2 Modelling the Covariance Structure for Multidimensional Abilities

For a test measuring several different proficiency dimensions, it is assumed that each examinee's proficiency follows a  $p$ -variate normal distribution with mean vector  $\boldsymbol{\mu}$  and the variance-covariance matrix  $\boldsymbol{\Sigma}$ . That is,  $\boldsymbol{\theta}_j \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ,  $\forall j = 1, 2, \dots, N$ . Since there is not much meaning in comparing abilities across dimensions, the mean of each dimension proficiency is set to zero. Thus, the mean vector for proficiency is set to a  $p$ -component zero vector.

Modelling the covariance matrix is very important but difficult because (1) there are  $\frac{p(p-1)}{2}$  parameters to estimate, where  $p$  is the number of dimensions; and (2) the matrix is required to be non-negative definite. To estimate the variance-covariance matrix  $\boldsymbol{\Sigma}$ , this study will use the inverse-Wishart ( $W^{-1}$ ) distribution, a multivariate generalization of the scaled inverse- $\chi^2$  distribution, as the prior distribution of the matrix  $\boldsymbol{\Sigma}$ , i.e.,

$$\boldsymbol{\Sigma} \sim W^{-1}(m, \boldsymbol{\Psi}), \quad (2.8)$$

which is suggested by Gelman, Carlin, Stern, and Rubin (2004). The above distribution is the conjugate prior distribution for the covariance matrix in a multivariate normal distribution. Where  $m$  and  $\boldsymbol{\Psi}$  describe the degrees of freedom and the scale

matrix for the inverse-Wishart distribution on  $\Sigma$ . The advantage of using inverse-Wishart as prior distribution for  $\Sigma$  is that the posterior distribution of  $\Sigma$  also follows the  $W^{-1}$  distribution (e.g., Gelman, Carlin, Stern, and Rubin, 2004) :

$$\Sigma \mid \boldsymbol{\theta} \sim W^{-1}(m + n, (n - 1)\mathbf{S} + \Psi + \frac{n\tau}{n + \tau}\bar{\boldsymbol{\theta}}\bar{\boldsymbol{\theta}}'), \quad (2.9)$$

where  $n$  is the number of examinees,  $\mathbf{S}$  is the sum of squares and cross product matrix about the sample mean

$$(n - 1)\mathbf{S} = \sum_{j=1}^N \boldsymbol{\theta}_j \boldsymbol{\theta}_j'. \quad (2.10)$$

$\tau$  is the number of prior measurements,  $\boldsymbol{\theta}_j$  is a  $p$  - dimension vector, and  $\bar{\boldsymbol{\theta}}$  is a  $p$ -dimensional sample mean vector. Since the posterior distribution on  $\Sigma$  is a known distribution,  $\Sigma \mid \boldsymbol{\theta}$  can be sampled directly.

Let  $\Sigma_k$  be the  $k$ th sample covariance matrix drawn from  $W^{-1}(m + n, (n - 1)\mathbf{S} + \Psi + \frac{n\tau}{n + \tau}\bar{\boldsymbol{\theta}}\bar{\boldsymbol{\theta}}')$ . Let  $s_{ijk}^2$  be the  $(ij)$ th component of  $\Sigma_k$ . Then the estimate of proficiency structure is the average of drawn covariance matrix samples:

$$\hat{\sigma}_{ijk}^2 = \frac{1}{N} \sum_{k=1}^N s_{ijk}^2, \quad (2.11)$$

where  $N$  is the total number of randomly drawn samples;  $i, j = 1, 2, \dots, p$ .

There are alternative approaches to modelling the underlying proficiency structure. Another method for estimating proficiency structure is addressed through a two-dimensional example. For a two-dimensional IRT model, assume the proficiency parameters come from a bivariate normal distribution  $N_2(\mathbf{0}, \Sigma)$ , where  $\Sigma$  is the stan-

standardized covariance matrix or correlation matrix, i.e.,

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Assume  $\rho$  has a prior density which is the uniform distribution on  $(-1, 1)$ . Then the posterior for  $\rho$ ,  $f_\rho(\rho \mid \Theta)$  can be expressed as

$$f_\rho(\rho \mid \Theta) \propto p(\Theta \mid \rho) I_{(-1, 1)}, \quad (2.12)$$

where  $p(\Theta \mid \rho)$  is the probability function given by

$$\begin{aligned} p(\Theta \mid \rho) &= \prod_{j=1}^N \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left[-\frac{1}{2(1-\rho^2)}(\theta_{1j}^2 - 2\rho\theta_{1j}\theta_{2j} + \theta_{2j}^2)\right] \\ &\propto (1-\rho^2)^{-\frac{N}{2}} \exp\left[-\frac{1}{2(1-\rho^2)} \sum_{j=1}^N (\theta_{1j}^2 - 2\rho\theta_{1j}\theta_{2j} + \theta_{2j}^2)\right], \end{aligned}$$

and  $I_{(-1, 1)}$  is an range indicator function.

Therefore, the posterior for  $\rho$  is

$$f_\rho(\rho \mid \Theta) \propto (1-\rho^2)^{-\frac{N}{2}} \exp\left[-\frac{1}{2(1-\rho^2)} \sum_{j=1}^N (\theta_{1j}^2 - 2\rho\theta_{1j}\theta_{2j} + \theta_{2j}^2)\right] I_{(-1, 1)}. \quad (2.13)$$

Let  $\rho = \frac{e^\xi - 1}{e^\xi + 1}$ . Then  $\xi = \log \frac{1+\rho}{1-\rho}$ .

$$\begin{aligned} f_\xi(\xi \mid \Theta) &= f_\rho(\rho \mid \Theta) \frac{d\rho}{d\xi} \\ &= f_\rho(\rho \mid \Theta) \frac{2e^\xi}{1+e^\xi}. \end{aligned}$$

Suppose  $\hat{\xi}$  is the maximum likelihood estimates of  $\xi$ , and  $\hat{\sigma}_\xi^2$  represents the estimated

variance of  $\hat{\xi}$ .  $\hat{\xi}$  can be obtained by letting

$$\begin{aligned}\hat{\rho} &= \arg \max_{\rho} p(\mathbf{\Theta} \mid \rho) \\ &= \arg \max_{\rho} \log p(\mathbf{\Theta} \mid \rho),\end{aligned}$$

where

$$\log p(\mathbf{\Theta} \mid \rho) = e + \frac{N}{2} \log(1 - \rho^2) - \frac{1}{2(1 - \rho^2)} \sum_{j=1}^N (\theta_{1j}^2 - 2\rho\theta_{1j}\theta_{2j} + \theta_{2j}^2),$$

where  $e$  is a constant. Solve the likelihood equation

$$\frac{\partial \log p(\mathbf{\Theta} \mid \rho)}{\partial \rho} = 0.$$

The equation above implies,

$$-\frac{N\rho}{1 - \rho^2} - \frac{\rho}{1 - \rho^2} \sum_{j=1}^N (\theta_{1j}^2 - 2\rho\theta_{1j}\theta_{2j} + \theta_{2j}^2) + \frac{1}{1 - \rho^2} \sum_{j=1}^N \theta_{1j}\theta_{2j} = 0,$$

i.e.,  $\hat{\rho}$  subjects to

$$N\rho + \frac{\rho}{1 - \rho^2} \sum_{j=1}^N (\theta_{1j}^2 - 2\rho\theta_{1j}\theta_{2j} + \theta_{2j}^2) - \sum_{j=1}^N \theta_{1j}\theta_{2j} = 0.$$

Note here, the prior  $\pi_{\rho}(\rho) = U(-1,1)$ . So  $\hat{\rho}_{mle} = \tilde{\rho}_{mode}$ . Thus  $\hat{\xi} = \log \frac{1 + \hat{\rho}}{1 - \hat{\rho}}$ . The

Fisher information

$$\begin{aligned}I(\rho) &= \text{var}\left(\frac{\partial}{\partial \rho} \log p(\mathbf{\Theta} \mid \rho)\right) \\ &= -E \frac{\partial^2}{\partial \rho^2} \log p(\mathbf{\Theta} \mid \rho) = \frac{N}{1 - \rho^2}.\end{aligned}$$

Then the asymptotic distribution of  $\hat{\rho}_{mle}$  is approximated by  $N(\rho, \frac{1}{NI(\rho)})$  as  $N \rightarrow$

$\infty$ .

Therefore, by the delta method,  $\hat{\xi}$  has the asymptotic distribution as  $\hat{\xi} \rightarrow N(h(\rho), \frac{1}{NI(\rho)}h'(\rho))$ , where  $h(\rho) = \log \frac{1+\rho}{1-\rho} = \xi$ ,  $h'(\rho) = \frac{2}{1-\rho^2}$ , i.e.,  $\hat{\xi} \sim N(\xi, \frac{e^\xi + 1}{2Ne^\xi})$ ,  $\hat{\sigma}_\xi^2 = \frac{e^\xi + 1}{2Ne^\xi}$ . Hence a Metropolis-Hastings algorithm can be written to generate  $\xi$  from  $f_\xi(\xi | \Theta)$  using  $N(\hat{\xi}, \hat{\sigma}_\xi^2)$  as the proposal density.

Since the target function  $f_\xi(\xi | \Theta) \rightarrow N(\hat{\xi}, \hat{\sigma}^2)$ . The sampling density is  $N(\hat{\xi}, \hat{\sigma}^2)$ . The transition function can be expressed as  $r_\xi(\cdot) = N(\hat{\xi}, \hat{\sigma}^2)$ . The M-H algorithm is as follows: Given  $\xi^t$ , simulate  $y$  from  $N(\hat{\xi}, \hat{\sigma}^2)$ , then  $\xi^{t+1} = y$  with  $\alpha(\xi^t, y)$  and  $\xi^t$  with  $1 - \alpha(\xi^t, y)$ , where

$$\alpha(\xi, y) = \min \left\{ \frac{f_\xi(y | \Theta) \phi(\frac{\xi - \hat{\xi}}{\hat{\sigma}_\xi})}{f_\xi(\xi | \Theta) \phi(\frac{y - \hat{\xi}}{\hat{\sigma}_\xi})}, 1 \right\}.$$

Repeat this step.

### 2.3.3 Random Walk Metropolis Algorithm within Gibbs

Since each complete conditional distribution is not convenient for sampling directly from the expressions given in Section 2.3.1, a Metropolis step, in which each parameter or block has to specify a proposal distribution, is needed for the sampling process. Patz and Junker (1997) point out that there is much freedom in choosing the proposal distributions. For example, to sample a proposal value for  $\theta_j$  at step  $t + 1$ , a multivariate normal distribution can be chosen as the convenient proposal distribution.

The random walk algorithm will choose the candidate state via a random walk mechanism. The candidate state is not chosen independently of the current state. And

the candidate state is not always accepted, unlike in the Gibbs sampler. Specifically, let  $\mathbb{R}^p$  be the  $p$ -dimensional Euclidian space, and let  $r$  be a density on  $\mathbb{R}^p$  so that the transition function is defined as  $R(y, B) = \int_B r(z - y)dz$ . Define the acceptance probability  $\alpha$  by

$$\alpha(y, z) = \min \left\{ \frac{g(z)r(y - z)}{g(y)r(z - y)}, 1 \right\}, \quad (2.14)$$

where  $g(\cdot)$  is the density of the target distribution function (e.g., the above posteriors for each examinee proficiency parameter,  $P_\theta(\theta_j \mid \Theta_{-j}, \Sigma, \mathbf{A}, \mathbf{d}, \mathbf{c}, \mathbf{U})$ , or the complete conditional distribution for each item parameters,  $P_{\mathbf{a}}(\mathbf{a}_i \mid \mathbf{A}_{-i}, \Theta, \Sigma, \mathbf{d}, \mathbf{c}, \mathbf{U})$ ,  $P_d(d_i \mid \mathbf{d}_{-i}, \Theta, \Sigma, \mathbf{A}, \mathbf{c}, \mathbf{U})$ ,

and  $P_c(c_i \mid \mathbf{c}_{-i}, \Theta, \Sigma, \mathbf{A}, \mathbf{d}, \mathbf{U})$ ). If the denominator is zero, just set  $\alpha = 1$ . Suppose  $Y_t = y$ . Generate a “candidate” observation  $z$  from the distribution  $R(y, \cdot)$ ; accept this observation (set  $Y_{t+1} = z$ ) with probability  $\alpha(y, z)$ . Otherwise, reject this observation (set  $Y_{t+1} = Y_t = y$ ). Another way to describe the procedure is as follows. Start at  $y$ . Generate a candidate step  $w$  from the distribution  $R$  defined by  $R(B) = \int_B r(x)dx$  with probability  $\alpha(y, y + w)$  moving forward to  $w$ ; Otherwise stay at  $y$ .

In the MIRT context, for instance, denote  $r_\theta(\theta_j^t, \theta_j^{t+1})$  as the transition function for the constructed Markov chain for sampling the  $j$ th examinee’s abilities. For random walk Metropolis algorithm, the transition kernel can have the form

$$r_\theta(\theta_j^t, \theta_j^{t+1}) = \exp \left\{ -\frac{1}{2}(\theta_j^t - \theta_j^{t+1})' \Sigma^{-1}(\theta_j^t - \theta_j^{t+1}) \right\}. \quad (2.15)$$

Then the acceptance probability for the new candidate  $\theta_j^*, j = 1, 2, \dots, N$  from the



transition kernel  $r_{\theta}(\theta_j^t, \theta_j^{t+1})$  is

$$\alpha(\theta_j^t, \theta_j^*) = \min \left\{ \frac{g_{\theta}(\theta_j^*) r_{\theta}(\theta_j^*, \theta_j^t)}{g_{\theta}(\theta_j^t) r_{\theta}(\theta_j^t, \theta_j^*)}, 1 \right\}. \quad (2.16)$$

Note here the target distribution  $g_{\theta}(\cdot)$  is the complete conditional distribution defined previously, i.e.,

$$\begin{aligned} g_{\theta}(\theta_j) &= P_{\theta}(\theta_j \mid \Theta_{-j}, \mathbf{A}, \mathbf{d}, \mathbf{c}, \mathbf{U}) \propto \mathbf{L}(\mathbf{u}_j \mid \theta_j, \mathbf{A}, \mathbf{d}, \mathbf{c}) \pi_{\theta}(\theta_j \mid \Sigma) \\ &= \prod_{i=1}^n p_i(\theta_j)^{u_{ij}} (1 - p_i(\theta_j))^{1 - u_{ij}} \pi_{\theta}(\theta_j). \end{aligned}$$

Similarly, the acceptance probability for a new candidate of item parameters  $\mathbf{a}_i^*$  for

item  $i$ ,  $i = 1, 2, \dots, n$  from the transition kernel  $r_{\mathbf{a}}(\mathbf{a}_i^t, \mathbf{a}_i^{(t+1)})$  is,

$$\alpha(\mathbf{a}_i^t, \mathbf{a}_i^*) = \min \left\{ \frac{g_{\mathbf{a}}(\mathbf{a}_i^*) r_{\mathbf{a}}(\mathbf{a}_i^*, \mathbf{a}_i^t)}{g_{\mathbf{a}}(\mathbf{a}_i^t) r_{\mathbf{a}}(\mathbf{a}_i^t, \mathbf{a}_i^*)}, 1 \right\}, \quad (2.17)$$

where  $g_{\mathbf{a}}(\cdot)$  is the complete conditional distribution for  $\mathbf{a}_i$ , i.e.,  $g_{\mathbf{a}}(\mathbf{a}_i^*) \propto L(\mathbf{v}_i \mid \Theta, \mathbf{a}_i^*, d_i, c_i) \pi_{\mathbf{a}}(\mathbf{a}_i^*)$ . In the same way, we can find  $\alpha(d_i^t, d_i^*)$  and  $\alpha(c_i^t, c_i^*)$ .

The following are the proposal densities corresponding to person and item parameters, which are chosen for the purpose of convenience and efficiency.

Proposal density for  $\theta^{t+1}$  is  $N_p(\theta^t, \Sigma_{\theta^t})$ .

Proposal density for each component of  $\mathbf{a}_i^{t+1}$ ,  $a_{ik}$  is  $U(a_{ik}^t - h, a_{ik}^t + h)$ ,

$k = 1, 2, \dots, p$ ;

Proposal density for  $d_i^{t+1}$  is  $N(d_i^t, \sigma^2)$ .

Proposal density for  $c_i^{t+1}$  is  $U(c_i^t - \delta, c_i^t + \delta)$ ;

where  $h$ ,  $\delta$ , and  $\sigma^2$  are constants. In this study,  $h = 0.3$ ,  $\delta = .03$ , and  $\sigma^2 = 1$ .

Once the derivation of the complete conditional distribution for each parameter in the multidimensional model is finished, the corresponding acceptance probabilities can be calculated. And if the proposal densities are specified, it is ready to draw parameter samples.

The steps for this drawing of parameter samples for the MIRT model are:

1. Draw  $\theta_j^* \sim N_p(\theta_j^t, \Sigma_{\theta^t})$ ,  $\forall j = 1, 2, \dots, N$ .  $\theta_j^{t+1} = \theta_j^*$  has acceptance probability  $\alpha(\theta_j^t, \theta_j^*)$
2. Draw  $\Sigma \mid \theta \sim W^{-1}(m + n, (n - 1)S + \Psi + \frac{n\tau}{n + \tau}\bar{\theta}\bar{\theta}')$
3. Draw each  $a_{ik}^* \sim U(a_{ik}^t - h, a_{ik}^t + h)$ ,  $a_{ik}^{t+1} = a_{ik}^*$  with probability of  $\alpha(a_{ik}^t, a_{ik}^*) \forall k = 1, 2, \dots, p$  and  $i = 1, 2, \dots, n$ .  $p$  is the total number of dimensions.
4. Draw  $d_i^* \sim N(d_i^t, \sigma^2)$  with acceptance probability of  $\alpha(d_i^t, d_i^*) \forall i = 1, 2, \dots, n$ .
5. Draw  $c_i^* \sim U(c_i^t - \delta, c_i^t + \delta)$  with acceptance probability of  $\alpha(c_i^t, c_i^*) \forall i = 1, 2, \dots, n$ . Here  $h$  and  $k$  are known constants.

## 2.4 Unbiased and Consistent Estimators of Parameters

Let  $\hat{\theta}_{jk}$ ,  $\hat{a}_{ik}$ ,  $\hat{d}_i$ ,  $\hat{c}_i$  be the model estimators  $\forall j = 1, 2, \dots, N$ ,  $i = 1, 2, \dots, n$ ,  $k = 1, 2, \dots, p$ . For example, if the samples from the complete conditional distribution of  $\theta_j$ ,  $a_i$ ,  $d_i$ ,  $c_i$  are drawn from the constructed Markov chain, then  $\hat{\theta}_{jk} = \frac{1}{M} \sum_{m=1}^M \theta_{jk}^m$ ,

$\hat{a}_{ik} = \frac{1}{M} \sum_{m=1}^M a_{ik}^m$ ,  $\hat{d}_i = \frac{1}{M} \sum_{m=1}^M d_i^m$ , and  $\hat{c}_i = \frac{1}{M} \sum_{m=1}^M c_i^m$ , where  $M$  is the sample size used for the estimates after certain length of the burn-in period.

Obviously,  $E(\hat{\theta}_j) = \theta_j$ ,  $E(\hat{\mathbf{a}}_i) = \mathbf{a}_i$ ,  $E(\hat{d}_i) = d_i$ ,  $E(\hat{c}_i) = c_i$  since  $E\theta_{jk}^m = \theta_j$ ,  $Ea_{ik}^m = a_{ik}$ ,  $Ed_i^m = d_i$ , and  $Ec_i^m = c_i$ . That is, the estimators are unbiased.

The variance of the estimates  $Var(\hat{\theta}_{jk}) = \frac{Var(\theta_{jk})}{M} \rightarrow 0, \forall k = 1, 2, \dots, p$ , as  $M \rightarrow \infty$ .  $Var(\hat{a}_{ik}) = \frac{Var(a_{ik})}{M} \rightarrow 0$ ,  $Var(\hat{d}_i) = \frac{Var(d_i)}{M} \rightarrow 0$ ,  $Var(\hat{c}_i) = \frac{Var(c_i)}{M} \rightarrow 0$ , as  $M \rightarrow \infty$ . By the law of large number,  $\hat{\theta}_j \rightarrow \theta_j$ ,  $\hat{\mathbf{a}}_i \rightarrow \mathbf{a}_i$ ,  $\hat{d}_i \rightarrow d_i$ , and  $\hat{c}_i \rightarrow c_i$  in probability. Therefore, the estimates are consistent.

By the central limit theorem,

$$\frac{\hat{\theta}_{jk} - \theta_{jk}}{\sqrt{var(\hat{\theta}_{jk})}} \Rightarrow N(0, 1), \quad (2.18)$$

as  $M \rightarrow \infty$ , for  $j = 1, 2, \dots, N$ . This can give a confidence interval for the estimate of proficiency parameters. Similarly, the results also hold for item parameter estimates.

## Chapter 3

# Simulation Studies and Results

The derivations for the application of MCMC methods into the 3-PL linear logistic multidimensional IRT model are illustrated in Chapter 2. This approach is implemented in a C++ program, which provides an efficient computational tool for parameter estimation of MIRT models of the application of the program are reported in the chapter. In this chapter, the parameter estimates for MIRT models. The accuracy and stability of the MCMC estimates will be examined by simulating various testing situations for the one-, three-, and five-dimensional MIRT models, respectively.

Various simulation studies are presented in this chapter in an attempt to examine the effects of four potential factors on the recovery of item and underlying proficiency parameters. These factors are: the number of proficiency dimensions, proficiency structure (i.e., covariance matrix for the proficiency distribution), test length (i.e., the number of test items), and the sample size (i.e., the number of examinees). Using simulated data to investigate parameter estimation has at least two advantages:

(1) since the true person and item parameters are available, they can be used to assess the accuracy of parameter estimates, with smaller root mean square errors (RMSE) between the true parameters and the parameter estimates indicating more accurate estimation; (2) the information for the number of dimensions is available from the simulated data, as is similar to the confirmatory factor analysis given the factor structure is known before analyzing data. With knowing the number of dimensions, researchers do not have to do additional analysis to determine how many dimensions each item measures and what these dimensions are about, a strategy that can help researcher separate dimensionality analysis with the issue of parameter estimation. It is necessary to point out that determining the statistical dimension based on the observed data itself is actually a complex and active research area. For example, Researchers suggest detecting the underlying dimension structure by parametric approach (e.g., Reckase, Ackerman, & Carlson, 1988; Miller & Hirsh 1992) and nonparametric approach (e.g., Roussos, 1995). The topic of detecting dimension structure from the observed data is out of the scope of this research. Therefore, to control the dimensional structure in the simulated data instead of diagnosing it will facilitate an effective examination of the MCMC estimation approach.

In addition, to examine the performance of parameter estimation by the MCMC approach in this research involves only simulation experiments because: (1) real data analysis will bring the model-data fit issue, which is often confounded with the issue of parameter estimation and obviously is not the focus of this study; (2) it is more difficult to evaluate the accuracy of estimation due to the lack of the true parameter

information.

### 3.1 Prior Distributions for Model Parameters

The MCMC approach for parameter estimation is in fact from Bayesian perspective. The item and proficiency parameters are not treated as fixed values but random variables with probability distributions. The role of prior distributions for both item and proficiency parameters is to provide additional information on the parameters before data collection and parameter estimation. In this study, the prior distribution for proficiency vector is  $N_p(\mathbf{0}, \Sigma_\theta)$ . That is, the group of examinees is assumed to come from the multivariate normal population  $N_p(\mathbf{0}, \Sigma_\theta)$ , where  $p$  is the number of dimensions. The prior distribution for each component of each a parameter is the uniform distribution, the prior distribution for each  $d$  parameter is the standard normal distribution, and the prior for each  $c$  parameter is also the uniform distribution.

### 3.2 Diagnosing the Convergence of Markov Chains

There are many approaches to the diagnosis of the convergence of a Markov chain. The purpose of this analysis is to ensure that the constructed Markov chains for the posterior distributions for both item and proficiency parameters through the Metropolis-Hastings within Gibbs algorithm have the target stationary distributions before taking sample for Monte Carlo estimation. The reliable estimation requires that each posterior distribution of a parameter converges to its stationary distribution.

Gelfand and Smith (1990) suggested several approaches to check the convergence

based on graphical techniques. For  $m$  parallel chains, plot a histogram for  $n$  values of  $k$ th iteration, after skipping certain iterations (say  $p$  iterations), and plot a histogram for  $n$  values of  $(k + p)$ th iteration. Convergence is assumed if the histograms have very close pattern.

Gelman, Carlin, Stern, and Rubin (p.294, 2004) recommended an approach to the inference and assessing convergence based on several independent parallel chains. First, simulate several independent sequences, with over-dispersed initial values. If multiple chains with different starting values are well mixed after certain number of iterations, then one can conclude that the chain reaches the convergence.

### 3.3 Initial Values and Iterations

The choice of initial values should not affect the item and proficiency estimates, because the final estimates rely on the sample from the posterior distributions for the parameters when they reach stationary status. The initial values are often discarded before computing Monte Carlo estimates for the parameters. However, the initial values may affect the convergence speed for each chain of a posterior distribution. Thus, carefully selected starting values will accelerate the convergence speed and construct an effective Markov chain. For example, Beguin and Glass (1998) suggested using  $a = 1$ ,  $d = 0$ , and the true  $c$  parameter or its estimates from BILOG as starting values and concluded that 1000 burn-in iterations was sufficient.

In this study, random initial values will be used each time for the estimation. To

ensure the convergence of each chain, a large number of iterations, for example, 10,000, will be taken. Moreover, multiple chains (e.g., 3 chains) will be constructed for each data set to assess the convergence of each chain and evaluate the accuracy and stability of the estimates by comparing the estimation from each chain with different random initial values. Hence, the starting values used for estimating proficiency parameters in this study are randomly drawn from  $N_p(\mathbf{0}, \mathbf{I})$ , and the initial values for item parameters will be randomly sampled from uniform distributions.

Since three independent replications of Markov chains are constructed with different initial values for each data set, the final estimates for the parameters take the mean of the estimates from the three independent chains. For each independent chain, parameter estimates  $\tilde{H}$  is the average of the sample from posterior distributions, i.e.,

$$\tilde{H} = \frac{1}{n} \sum_{i=1}^n h_i, \quad (3.1)$$

where  $n$  is the number of samples drawn from the stationary Markov chain for the posterior distribution. Thus the final estimates of parameters for each data set  $\hat{H}$  is the average of the estimates from multiple independent chains,

$$\hat{H} = \frac{1}{m} \sum_{j=1}^m \tilde{H}_j, \quad (3.2)$$

where  $m$  is the number of replications, i.e.,  $m = 3$  in this study.

All of the data sets are randomly sampled from the linear logistic multidimensional IRT model for various conditions (e.g., test length, the sample size of examinees, the number of dimensions, and different proficiency covariance matrices). To minimize the sampling effects on parameter estimation, three replications are simulated for each



condition. Four factors considered in the simulation studies result in a total of 60 dichotomous response data sets. Therefore, the precision of parameter estimates can be compared across the sample size, the test length, and the proficiency structures.

### **3.4 Estimating the Unidimensional 3PL Model**

The form for the unidimensional 3PL model is given in equation (1.1) in the first section of Chapter 1. This section will discuss the parameter estimation by simulating dichotomous response data from the unidimensional 3PL model. One big difference for estimating unidimensional model parameters from the estimation of the multidimensional model parameters is that no underlying proficiency dimension structure needs to be estimated. To consider the model indeterminacy problem and establish a fixed metric for both item and proficiency parameter estimates, the sample of the posterior distributions for proficiency parameters will be standardized at each step of sample draw. Therefore, the final metric for the proficiency parameter estimates is placed on 0, 1 metric.

For the simulation study in this section, the underlying proficiency parameters and difficulty item parameters are generated from the standard normal distribution  $N(0, 1)$ ; the discriminating power and asymptote item parameters are generated from a uniform distribution. Two tests with 30 and 45 items were simulated. Each test is administrated to 2000 and 5000 examinees, respectively. The combination of the test length, the sample size, and replications yields 12 (i.e.,  $2 \times 2 \times 3$ ) data sets. Table

3.1 and Table 3.2 are the true items parameters for the two tests.

It can be seen from Table 3.1 and Table 3.2 that both tests contains a wide variety of values of item parameters. For example, in the 30-item test, the discriminating power  $a$  parameter ranges from the smallest of .54 to the largest of 2.43, the difficulty parameters from -1.64 to 1.6, and the asymptote parameters from 0 to .25. In the 45-item test, the discriminating parameters cover a range between .5 and 2.45, the difficulty parameters fall into a range within -1.78 to 2.85, and the asymptote parameters ranges from 0 to .25.

### 3.4.1 Assessing Convergence

Table 3.3 shows the three independent estimates from each chain replication with different initial values for the data set generated by the 30-item test to 2000 examinees. The final item parameter estimates are the mean of the three independent estimates for each chain. Clearly, the estimates from the three independent chains are very stable and consistent. For example, item 28 has the same estimates on  $a$  and  $c$  parameters over three chains, but has .01 difference on  $b$  parameter estimates across the three independent chains. The largest change for  $a$  parameter estimates over three independent chains is on item 1, showing 1.82 for the first chain, 1.67 for the second chain with a difference of .15, and 1.72 for the third chain. The slight change of estimates for each item parameter across the three independent chains indicates the stable estimates by the MCMC. More importantly, one can assess the convergence of the posterior distributions by the stability of the estimates over multiple chains

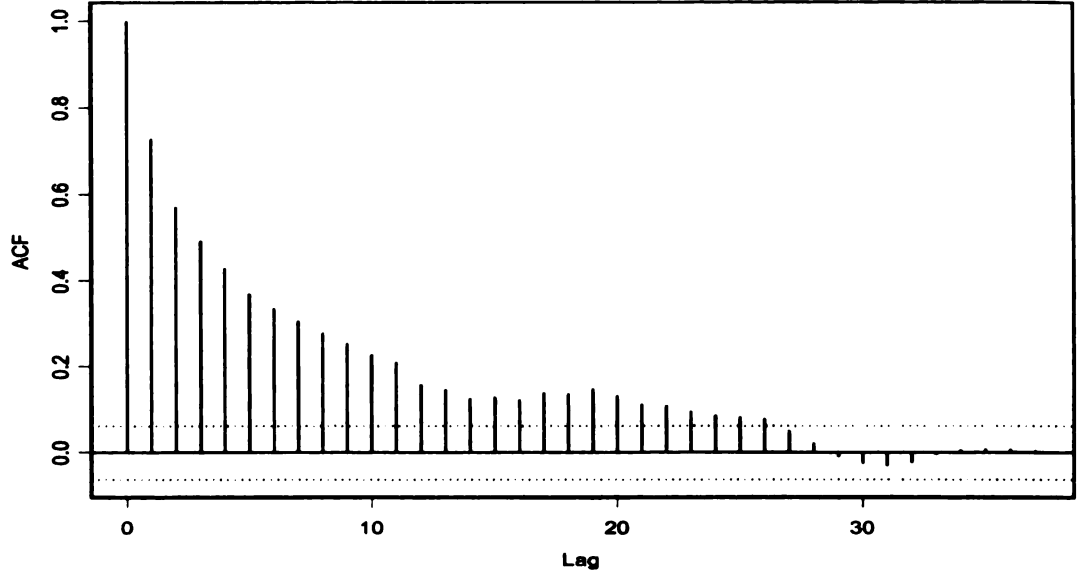
Table 3.1: True Item Parameters for 30-Item Test (Dim = 1)

Item	Discriminating ( $a$ )	Difficulty ( $b$ )	Asymptote ( $c$ )
1	1.67	-1.17	0.14
2	0.89	0.28	0.09
3	0.55	-1.64	0.14
4	1.85	-0.72	0.16
5	2.07	0.50	0.08
6	1.40	0.46	0.18
7	2.43	1.37	0.21
8	0.85	-0.04	0.14
9	1.39	0.91	0.09
10	1.25	0.14	0.09
11	1.52	-0.19	0.21
12	1.34	-0.80	0.25
13	1.64	-0.44	0.05
14	0.99	0.57	0.12
15	1.48	-1.11	0.16
16	0.54	0.48	0.09
17	1.78	1.60	0.03
18	1.10	0.21	0.14
19	2.09	-0.31	0.01
20	2.26	1.10	0.04
21	1.53	0.65	0.24
22	0.79	-0.41	0.11
23	2.40	0.57	0.11
24	0.73	-1.21	0.25
25	0.56	0.62	0.02
26	0.56	-1.43	0.19
27	1.01	1.51	0.04
28	2.07	1.31	0.18
29	2.05	-0.25	0.17
30	1.48	-1.62	0.10

Table 3.2: True Item Parameters for 45-Item Test (Dim = 1)

Item	$a$	$b$	$c$	Item	$a$	$b$	$c$
1	1.73	0.03	.09	24	1.98	-0.88	.09
2	2.45	0.00	.16	25	0.83	0.20	.02
3	2.35	0.45	.02	26	0.84	0.98	.07
4	1.04	0.15	.22	27	1.95	0.90	.01
5	2.37	0.27	.23	28	1.99	-0.51	.19
6	0.95	-1.78	.05	29	0.92	-1.80	.20
7	2.06	1.08	.08	30	1.26	2.85	.00
8	1.43	-0.59	.11	31	1.77	-1.19	.02
9	1.63	-0.67	.11	32	0.50	-0.44	.17
10	2.00	0.54	.18	33	1.78	-0.62	.08
11	2.13	0.33	.25	34	0.61	0.64	.00
12	1.27	-0.56	.17	35	2.21	-0.57	.19
13	1.45	-0.64	.22	36	2.31	0.54	.09
14	2.04	-1.31	.05	37	2.30	0.27	.07
15	0.53	1.16	.19	38	1.51	1.48	.07
16	1.51	-1.53	.13	39	2.26	0.45	.10
17	2.29	0.70	.10	40	0.85	-1.05	.14
18	0.62	-0.18	.07	41	1.33	-0.33	.15
19	2.10	-1.08	.11	42	1.02	-1.24	.02
20	1.69	0.64	.23	43	0.73	1.74	.11
21	1.55	-1.32	.08	44	1.21	1.53	.07
22	1.34	0.03	.18	45	1.99	1.31	.19
23	1.50	1.21	.03	-	-	-	-

Figure 3.1: Sample ACF for series of  $a_6$ , Dim = 1



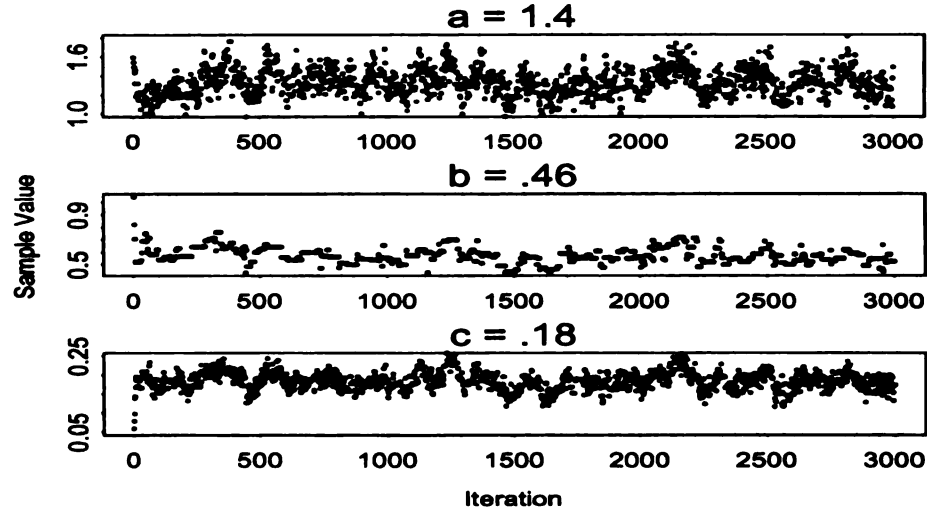
suggested by Gelman, Carlin, Stern, and Rubin (p.294, 2004). Table 3.3 provides numeric demonstrations that the chain has converged to its stationary distribution. Similar results are obtained for the sample size of 5000 and for the 45-item test but are omitted here.

Figure 3.1 describes the estimated autocorrelation function (ACF) in the series of discriminating power for the 5th item after throwing away the burn-in draws. It is found that the autocorrelation become negligible at lags greater than 28. Figure 3.2 illustrates the behavior of the Markov chains constructed by the M-H within Gibbs algorithm for item 5 in the 30-item test. The upper panel shows the first 2000 draws for the posterior distribution of  $a$  parameter, the middle shows the 2000 draws for the

Table 3.3: Estimates from three chains for 30-Item Test (Dim = 1, N = 2000)

Item	$a_1$	$a_2$	$a_3$	$d_1$	$d_2$	$d_3$	$c_1$	$c_2$	$c_3$
1	1.82	1.67	1.72	-1.00	-1.07	-1.05	.23	.18	.20
2	0.87	0.88	0.88	0.23	0.27	0.27	.07	.08	.08
3	0.46	0.45	0.43	-1.78	-1.80	-1.97	.20	.20	.17
4	1.66	1.58	1.60	-0.71	-0.75	-0.73	.16	.13	.14
5	2.09	2.12	2.12	0.56	0.58	0.57	.08	.09	.09
6	1.37	1.37	1.38	0.56	0.59	0.58	.20	.20	.20
7	2.09	2.11	2.10	1.43	1.44	1.44	.21	.21	.21
8	0.99	0.97	0.96	0.26	0.26	0.24	.26	.25	.25
9	1.50	1.49	1.51	0.90	0.91	0.90	.07	.07	.07
10	1.45	1.42	1.45	0.18	0.19	0.20	.08	.08	.09
11	1.70	1.71	1.71	-0.12	-0.10	-0.11	.22	.23	.23
12	1.17	1.13	1.17	-0.83	-0.87	-0.82	.23	.20	.24
13	1.68	1.68	1.63	-0.33	-0.31	-0.34	.06	.07	.05
14	1.01	1.00	0.99	0.64	0.65	0.64	.12	.12	.11
15	1.51	1.54	1.53	-1.16	-1.12	-1.14	.05	.07	.06
16	0.64	0.68	0.66	0.83	0.90	0.89	.18	.20	.19
17	1.82	1.80	1.81	1.68	1.70	1.69	.04	.04	.04
18	1.19	1.20	1.21	0.17	0.20	0.20	.12	.13	.13
19	2.06	2.06	2.05	-0.31	-0.29	-0.30	.00	.00	.00
20	2.35	2.35	2.36	1.17	1.19	1.18	.04	.04	.04
21	1.60	1.60	1.59	0.76	0.77	0.77	.23	.23	.22
22	0.79	0.8	0.75	-0.31	-0.26	-0.37	.14	.16	.12
23	2.24	2.24	2.22	0.60	0.63	0.62	.12	.12	.12
24	0.67	0.65	0.67	-1.52	-1.55	-1.50	.07	.06	.08
25	0.58	0.59	0.59	0.78	0.79	0.81	.04	.04	.04
26	0.59	0.57	0.60	-1.33	-1.37	-1.30	.20	.20	.22
27	1.14	1.18	1.14	1.45	1.45	1.46	.04	.04	.04
28	2.22	2.22	2.22	1.36	1.37	1.37	.19	.19	.19
29	2.34	2.31	2.34	-0.20	-0.20	-0.19	.18	.18	.18
30	1.72	1.72	1.69	-1.41	-1.39	-1.43	.23	.24	.22

Figure 3.2: Sample draw at first 3000 iterations for series of  $a$ ,  $b$  and  $c$



posterior distribution of  $b$  parameter, and the lower panel gives the first 2000 draws of the posterior distribution for the asymptote parameter. The path plot in Figure 3.2 shows that the posterior distributions for the fifth item parameters mixed well even in the first 2000 draws. The path plots for other items in the 30-item test or 45-item tests have similar path plots for 2000 draws and are not shown.

The column 2 through 7 (denoted as  $\hat{a}, \hat{b}, \hat{c}, S(a), S(b), S(c)$ ) in Table 3.4 are the item parameter estimates and their corresponding standard error of the estimates for the 30-item test with sample size 2000 from the first replication of response data. The last six columns are the values for the sample size 5000. Table 3.5 shows the item parameter estimates from BILOG-MG3 using MML procedure for the same

Table 3.4: Item Parameter Estimates for 30-Item Test (Dim = 1)

Item	N = 2000						N = 5000					
	$\hat{a}$	$\hat{b}$	$\hat{c}$	$S(a)$	$S(b)$	$S(c)$	$\hat{a}$	$\hat{b}$	$\hat{c}$	$S(a)$	$S(b)$	$S(c)$
1	1.74	-1.04	.20	.16	.07	.06	1.70	-1.14	.13	.10	.07	.05
2	0.88	0.26	.08	.09	.09	.04	0.83	0.22	.04	.04	.04	.02
3	0.45	-1.85	.19	.05	.29	.09	0.58	-1.35	.23	.04	.17	.06
4	1.61	-0.73	.14	.18	.10	.07	1.88	-0.65	.17	.12	.05	.03
5	2.11	0.57	.09	.19	.04	.01	2.07	0.55	.07	.11	.02	.01
6	1.37	0.58	.20	.15	.05	.02	1.41	0.56	.18	.09	.03	.01
7	2.10	1.44	.21	.27	.05	.01	2.17	1.42	.21	.20	.03	.03
8	0.97	0.25	.25	.09	.08	.03	0.75	-0.10	.09	.05	.07	.03
9	1.50	0.90	.07	.14	.04	.01	1.38	0.98	.08	.10	.02	.01
10	1.44	0.19	.08	.10	.04	.02	1.20	0.20	.09	.06	.03	.02
11	1.71	-0.11	.23	.15	.05	.03	1.64	-0.08	.24	.10	.04	.02
12	1.16	-0.84	.22	.10	.10	.06	1.29	-0.78	.22	.08	.07	.04
13	1.66	-0.33	.06	.12	.04	.03	1.55	-0.42	.01	.07	.02	.01
14	1.00	0.64	.12	.12	.08	.03	0.93	0.63	.12	.07	.06	.02
15	1.53	-1.14	.06	.13	.08	.06	1.62	-0.95	.18	.11	.06	.04
16	0.66	0.87	.19	.10	.13	.04	0.55	0.57	.11	.04	.08	.03
17	1.81	1.69	.04	.25	.07	.01	1.83	1.66	.03	.14	.03	.00
18	1.20	0.19	.13	.12	.07	.03	1.16	0.23	.13	.07	.04	.02
19	2.06	-0.30	.00	.13	.03	.01	2.04	-0.27	.01	.09	.01	.01
20	2.35	1.18	.04	.15	.04	.01	2.24	1.15	.05	.15	.02	.00
21	1.60	0.77	.23	.18	.05	.02	1.60	0.69	.24	.12	.02	.01
22	0.78	-0.31	.14	.06	.12	.05	0.71	-0.51	.04	.04	.09	.04
23	2.23	0.62	.12	.18	.04	.01	2.21	0.62	.10	.14	.02	.01
24	0.66	-1.52	.07	.04	.14	.06	0.70	-1.41	.10	.03	.10	.06
25	0.59	0.79	.04	.06	.10	.03	0.58	0.74	.03	.04	.09	.03
26	0.59	-1.33	.21	.06	.26	.09	0.55	-1.69	.04	.03	.11	.04
27	1.15	1.45	.04	.13	.07	.01	1.09	1.54	.05	.08	.04	.01
28	2.22	1.37	.19	.21	.05	.01	2.35	1.35	.18	.14	.03	.01
29	2.33	-0.20	.18	.16	.04	.02	2.14	-0.14	.20	.13	.02	.02
30	1.71	-1.41	.23	.17	.08	.06	1.79	-1.35	.23	.15	.08	.06



Table 3.5: Item Parameter Estimates for 30-Item Test In BILOG-MG3 (Dim = 1)

Item	N = 2000			N = 5000		
	$\hat{a}$	$\hat{b}$	$\hat{c}$	$\hat{a}$	$\hat{b}$	$\hat{c}$
1	1.83	-1.00	.25	1.71	-1.17	.15
2	0.88	0.21	.08	0.81	0.18	.04
3	0.61	-0.65	.50	0.58	-1.44	.21
4	1.67	-0.73	.16	1.90	-0.68	.18
5	2.09	0.52	.08	2.04	0.51	.07
6	1.39	0.54	.20	1.39	0.53	.18
7	2.20	1.38	.21	2.14	1.39	.21
8	0.95	0.17	.24	0.75	-0.12	.11
9	1.49	0.85	.07	1.38	0.94	.09
10	1.45	0.15	.08	1.17	0.14	.08
11	1.66	-0.17	.22	1.62	-0.13	.24
12	1.18	-0.83	.25	1.27	-0.83	.21
13	1.65	-0.37	.05	1.53	-0.47	.01
14	1.00	0.58	.11	0.93	0.60	.12
15	1.46	-1.23	.00	1.64	-0.97	.19
16	0.67	0.84	.19	0.57	0.58	.12
17	1.82	1.62	.04	1.81	1.63	.03
18	1.23	0.17	.13	1.15	0.18	.12
19	2.02	-0.33	.00	2.02	-0.31	.01
20	2.52	1.12	.04	2.21	1.12	.04
21	1.60	0.72	.23	1.57	0.66	.24
22	0.78	-0.35	.14	0.66	-0.61	.00
23	2.28	0.56	.12	2.19	0.59	.10
24	0.64	-1.67	.00	0.72	-1.31	.17
25	0.58	0.73	.03	0.55	0.64	.01
26	0.57	-1.52	.15	0.51	-1.87	.00
27	1.16	1.40	.04	1.08	1.51	.05
28	2.41	1.31	.19	2.47	1.31	.18
29	2.51	-0.21	.19	2.09	-0.19	.20
30	1.72	-1.39	.28	1.82	-1.36	.26

Table 3.6: Item Parameter Estimates for 45-Item Test (Dim = 1)

Item	N = 2000						N = 5000					
	$\hat{a}$	$\hat{b}$	$\hat{c}$	$S(a)$	$S(b)$	$S(c)$	$\hat{a}$	$\hat{b}$	$\hat{c}$	$S(a)$	$S(b)$	$S(c)$
1	1.54	-0.01	.08	.10	.03	.02	1.59	0.01	.07	.07	.03	.01
2	2.42	0.06	.17	.11	.03	.02	2.31	.08	.17	.13	.02	.01
3	2.31	0.46	.02	.15	.02	.01	2.41	0.50	.02	.10	.02	.00
4	1.00	0.20	.24	.10	.07	.03	0.95	0.20	.20	.05	.03	.02
5	2.37	0.31	.26	.14	.04	.02	2.44	0.35	.24	.09	.03	.01
6	0.96	-1.65	.09	.08	.14	.07	0.99	-1.55	.11	.05	.08	.05
7	2.10	1.13	.07	.20	.03	.01	2.06	1.16	.08	.16	.02	.01
8	1.35	-0.60	.14	.14	.11	.06	1.41	-0.49	.15	.07	.04	.03
9	1.55	-0.69	.11	.12	.06	.04	1.66	-0.62	.12	.08	.03	.02
10	1.93	0.55	.16	.18	.04	.02	2.08	0.58	.17	.12	.02	.01
11	1.87	0.33	.24	.17	.05	.02	2.03	0.37	.24	.12	.03	.01
12	1.39	-0.42	.23	.14	.10	.06	1.33	-0.47	.18	.08	.04	.03
13	1.28	-0.71	.18	.15	.12	.06	1.36	-0.63	.18	.06	.04	.03
14	2.29	-1.24	.08	.17	.06	.05	2.23	-1.19	.05	.12	.03	.03
15	0.50	1.01	.15	.08	.17	.05	0.43	1.02	.12	.05	.13	.04
16	1.53	-1.53	.21	.15	.11	.06	1.57	-1.43	.20	.18	.15	.11
17	1.89	0.76	.09	.17	.04	.01	2.16	0.75	.09	.12	.02	.01
18	0.61	-0.13	.10	.06	.16	.06	0.59	-0.25	.04	.03	.10	.04
19	2.28	-1.07	.14	.15	.05	.04	1.91	-1.07	.07	.13	.05	.03
20	1.62	0.70	.23	.17	.06	.02	1.77	0.71	.25	.13	.03	.01
21	1.60	-1.30	.09	.13	.08	.06	1.43	-1.35	.02	.07	.04	.02
22	1.38	0.04	.19	.12	.06	.03	1.45	0.11	.19	.07	.02	.01

Table 3.7: Item Parameter Estimates for 45-Item Test (Dim = 1), cont.

Item	N = 2000						N = 5000					
	$\hat{a}$	$\hat{b}$	$\hat{c}$	$S(a)$	$S(b)$	$S(c)$	$\hat{a}$	$\hat{b}$	$\hat{c}$	$S(a)$	$S(b)$	$S(c)$
23	1.40	1.25	.02	.12	.05	.01	1.51	1.25	.03	.08	.03	.00
24	1.84	-0.96	.04	.14	.05	.03	2.08	-0.78	.12	.09	.02	.02
25	.85	0.24	.02	.06	.06	.02	0.83	0.19	.00	.03	.03	.01
26	1.03	1.19	.12	.14	.07	.02	0.87	1.13	.09	.06	.04	.01
27	1.69	1.01	.01	.16	.04	.01	1.85	0.98	.01	.10	.02	.00
28	1.80	-0.52	.24	.15	.04	.03	1.95	-0.45	.21	.12	.05	.03
29	0.87	-1.95	.13	.10	.22	.10	0.85	-1.94	.07	.05	.09	.05
30	1.19	3.17	.00	.20	.29	.00	1.26	2.98	.00	.09	.10	.00
31	1.72	-1.22	.02	.11	.05	.03	1.69	-1.12	.04	.12	.07	.04
32	0.44	-0.72	.11	.04	.21	.06	0.50	-.67	.05	.02	.10	.04
33	1.66	-0.63	.06	.13	.05	.03	1.76	-0.56	.05	.08	.03	.02
34	0.64	0.69	.02	.06	.09	.02	0.66	0.72	.03	.05	.05	.02
35	2.22	-0.58	.20	.20	.06	.04	2.40	-0.48	.21	.11	.03	.02
36	2.39	0.55	.08	.12	.03	.01	2.40	0.58	.09	.11	.02	.01
37	2.37	0.32	.08	.14	.03	.01	2.46	0.33	.08	.07	.02	.01
38	1.37	1.56	.07	.19	.06	.01	1.47	1.54	.07	.11	.04	.01
39	2.33	0.50	.10	.14	.03	.01	2.37	0.49	.09	.11	.02	.01
40	0.76	-1.24	.05	.05	.09	.05	0.80	-1.13	.07	.04	.10	.05
41	1.18	-0.43	.06	.08	.05	.03	1.37	-0.24	.16	.06	.03	.02
42	1.32	-0.94	.21	.10	.08	.06	1.21	-1.01	.14	.05	.05	.04
43	0.77	1.76	.11	.13	.09	.02	0.77	1.78	.11	.08	.08	.01
44	1.19	1.59	.06	.13	.06	.01	1.27	1.60	.08	.09	.04	.01
45	1.84	1.36	.19	.24	.05	.01	1.94	1.34	.19	.16	.03	.01

data set, a standard procedure of item parameter estimation in most IRT calibration software. Comparing the results of item parameter estimates from these two different procedures, one can see that these results are very close to each other and close to their true item parameters, indicating the two estimation methods are comparable. Table 3.6 and 3.7 show the item parameter estimates and the corresponding standard error for the 45-item test.

As is true in many estimation programs in IRT, item parameter estimates contain estimation errors even if the data and the mathematical models have perfect fit. To examine the estimation accuracy of item parameter estimates, root mean square errors (RMSE) of the item parameter estimates are calculated from each data replication and each chain. In this study, three data replications are observed for both tests (i.e., the 30-item test and the 45-item test). Here data replication means, for example, the 30-item test is administered to three groups of different examinees who come from the same population ( $N(0, 1)$ ). Therefore, there will be three sets of item parameter estimates corresponding to the three groups of examinees. For each data set, the computation program will come up with three different chains along with three different initial values to make sure that the MCMC approach can provide stable parameter estimates. Each chain will independently give estimates for item parameters. Therefore, combining three data replications and three chains for each data set will yield nine sets of item parameter estimates. For each data set, the final item parameter estimates are the average of estimates from the three chains. RMSE is defined as the square root of the mean squared difference between the item parameter

Table 3.8: RMSE for Estimating Uni-dimensional Models (Dim = 1)

	30 × 2000	30 × 5000	45 × 2000	45 × 5000
<i>a</i>	.15	.07	.11	.07
<i>b</i>	.11	.08	.08	.08
<i>c</i>	.05	.04	.04	.03

estimates and the true item parameters over  $r$  data replications and across  $n$  items ( $r$  in this example is 3, and  $n$  is 30 or 45). Let  $\eta$  denote as item parameter (e.g., discriminating power parameter  $a$ , or difficulty parameter  $b$ , or asymptote parameter  $c$ ) and  $\hat{\eta}$  as item parameter estimates. Then RMSE can be calculated by

$$RMSE(\eta) = \sqrt{\frac{\sum_{i=1}^n \sum_{j=1}^r (\hat{\eta}_{ij} - \eta_{ij})^2}{r \times n}}.$$

RMSE gives a summary index of assessing the accuracy of item parameter estimates. Apparently, the larger RMSE of item parameter estimates for a data set, the worse of the item parameter estimates. For a simulation study, the perfect fit of model and data is assumed, and thus the difference between the true and item parameter estimates may depend on estimation procedures and some other factors (e.g., the sample size of examinees).

Table 3.8 contains the RMSE for item parameters. It shows that for the same test the larger the sample size, the smaller RMSE, and the less estimation errors. The largest RMSE for  $a$  is .15 in the 30-item test with 2000 examinees. The smallest RMSE is .07 in both tests when sample size is 5000. The largest RMSE for  $b$  is .11 in the 30-item test with examinee 2000. It also shows that the RMSE for  $c$  is generally smaller than RMSE for  $a$  and  $b$ , with the largest one .05 in the 30-item test to 2000

Table 3.9: Correlations Between True Proficiency and Estimates (Dim = 1)

Tests	$N = 2000$	$N = 5000$
30-items	.9546	.9554
45-items	.9712	.9718

examinees.

Table 3.9 shows the correlation between true proficiency and estimates from the MCMC approach. For the 30-item test, the correlations are around .96. The correlations in the 45-item test are about .97, slightly higher than those in the 30-item test. That is, longer tests gives higher correlation between true and estimates, implying better proficiency parameter estimation. Figure 3.3 shows the plots of true proficiency versus estimates corresponding to the four correlations in Table 3.9. One can see that the proficiency estimates from the longer test (i.e., the 45-item test) more closely around the reference line  $y = x$ , representing a higher correlation between the true and estimates. Figure 3.4 through Figure 3.6 are the plot of the true item parameter versus the estimates for parameter  $a$ ,  $b$ , and  $c$ , correspondingly. Most of the plots are close to the reference line  $y = x$ . For these figures that have larger sample size, the plots are more close to the reference line, implying better item parameter estimates.

### 3.5 Estimating the 3-Dimensional MIRT Model

This section will discuss the simulation studies of the parameter estimation for the 3-dimensional model, which is slightly different compared to the parameter estimation for the unidimensional model because the number of parameters in the multidimensional

Figure 3.3: True Proficiency Versus Estimates ( $Dim = 1$ )

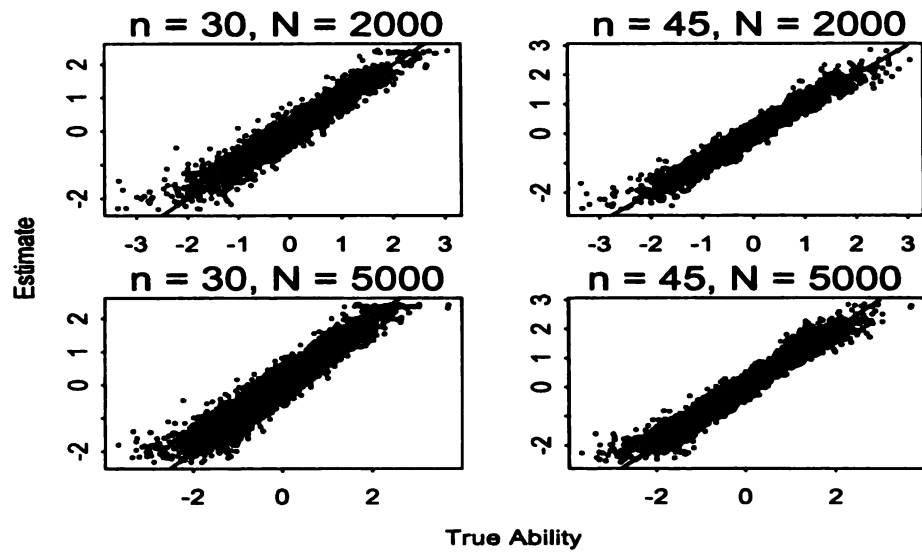


Figure 3.4: True  $a$  Parameter Versus Estimates ( $Dim = 1$ )

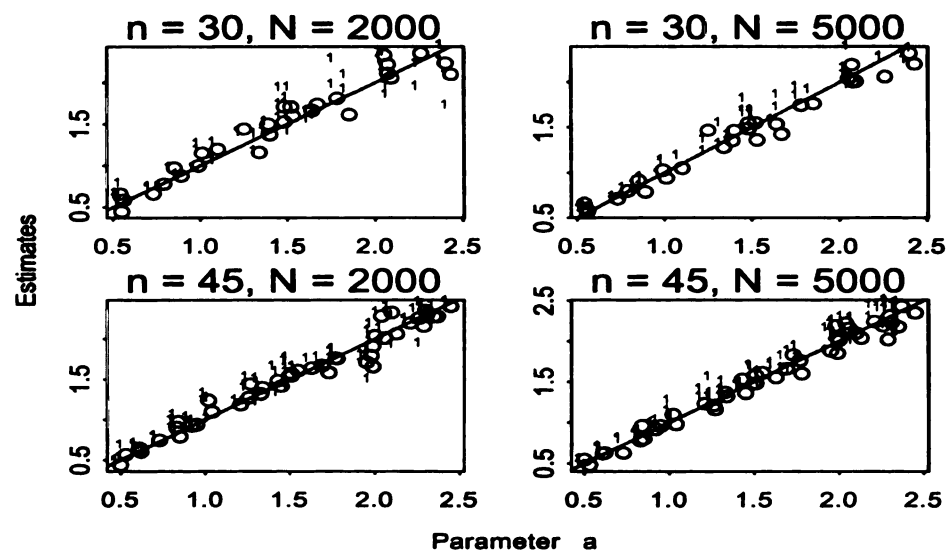


Figure 3.5: True  $b$  Parameter Versus Estimates ( $Dim = 1$ )

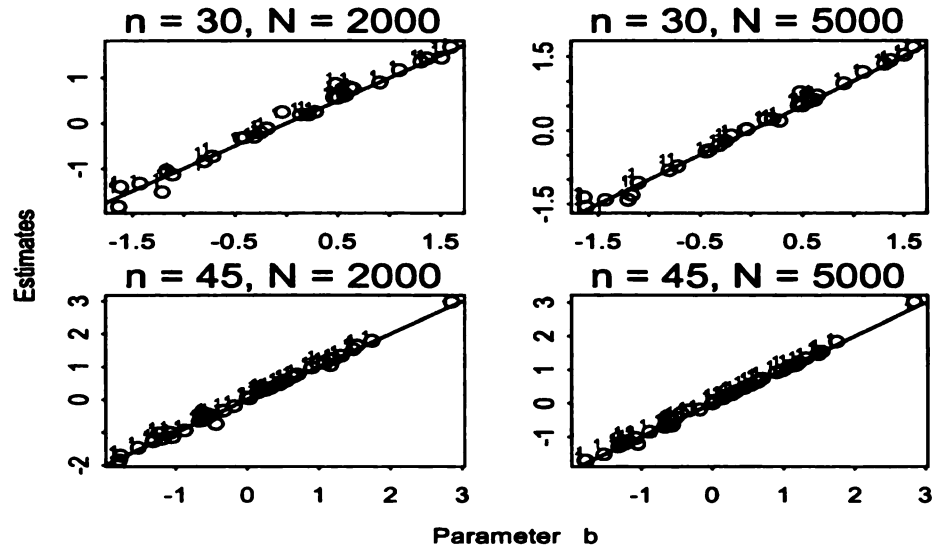
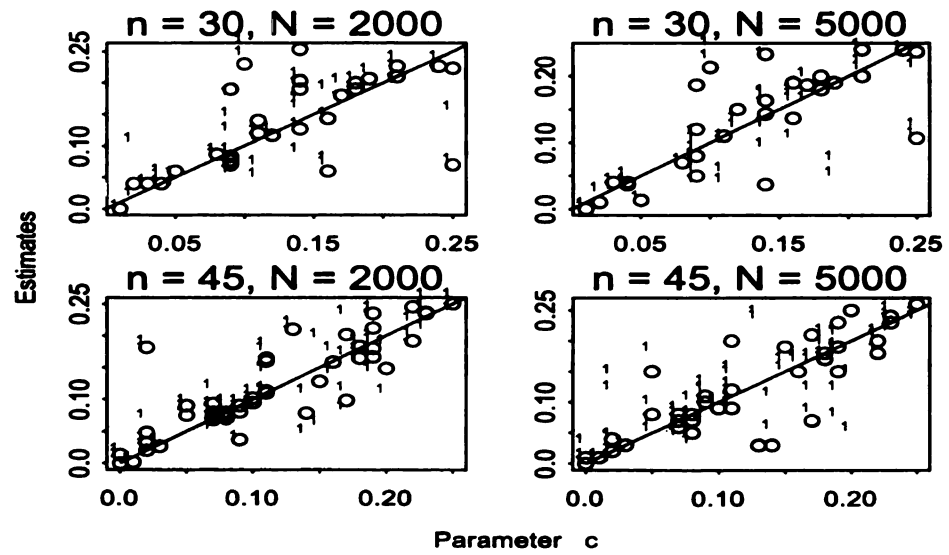


Figure 3.6: True  $c$  Parameter Versus Estimates ( $Dim = 1$ )





sional model is much greater than that in the unidimensional model. In addition, new parameters (e.g., proficiency structure parameters that appear as the components in the covariance matrix of the underlying proficiency distribution) need to be considered to estimate at the same time along with the estimation of the item and proficiency parameters. One more concern for MIRT model parameter estimation is the issue of indeterminacy that is inherited from the form of the MIRT model. Basically, one needs to put some constraints to ensure the MIRT model parameters have fixed solutions. The following sections will discuss the design of the simulation studies, for example, on how to generate the item and proficiency parameters and the response data, the underlying proficiency covariance, how to put constraints on the items in a test to establish a fixed scale for the parameter estimates, and how to assess the accuracy and stability for the parameter estimation.

### 3.5.1 Generating Proficiency Parameters

Assume that the underlying distribution of proficiency for each examinee follows the multivariate normal distribution with mean vector  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ . That is,  $\boldsymbol{\theta}_j \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ , where  $j = 1, 2, \dots, N$ . Proficiency parameters for each examinee are randomly drawn from  $N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ , where  $p$  is the number of dimensions;  $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$  is the generating covariance matrix, which corresponds to its dimensional structure and will have more discussions in Section 3.5.3. The mean vector  $\boldsymbol{\mu}$  here is set to  $\mathbf{0}$ , because each dimension actually represent one hypothetical construct and comparison among dimensions seems to be not necessary.

### 3.5.2 The Number of Proficiency Dimension and Sample Size

One factor that might indirectly affect the parameter estimation in the MIRT model is the proficiency dimensions (i.e., the number of latent variables in the complete latent space). As is known, the unidimensional IRT model ( $dim = 1$ ) has 3 parameters for each item and one parameter for an examinee's proficiency. For a test with  $n$  item and  $N$  examinees, the total number of parameters to be estimated is  $3n + N$ . But in the case of the 3-dimensional MIRT model, there are 5 parameters for each item (i.e., three  $a$  parameters plus  $d$  and  $c$  parameters), 3 parameters for an individual proficiency, and 3 more parameters for representing the components in the proficiency covariance matrix. Therefore, for a test with  $n$  items and  $N$  examinees, the total number of model parameters need to estimate is  $5n + 3N + 3$ , much more than that in the unidimensional model. The increasing number of parameters in the MIRT model brings more difficulties for the estimation given the test length  $n$  and the sample size of examinees  $N$ , since more information is required to achieve the same level of estimation precision.

The simulation studies here consider two different numbers of dimensions for estimating multi-dimensional MIRT models: three and five proficiency dimensions. That is, three-, and five-dimensions of proficiency are required to determine the correct answers in the simulation studies.

The stable Monte Carlo estimates may depend on the sample size (this would also be the case for the maximum likelihood and Bayesian modal estimation). To investigate the effect of the sample size on the accuracy and stability of the estimation, the

response data with the sample size 2000 and 5000 examinees are independently generated from the multivariate normal population. The sample size 2000 is considered as moderate, and 5000 as a large sample.

### 3.5.3 Proficiency Structure

For multivariate analysis, the estimating of the covariance matrix is an important step, because the covariance structure can reveal some helpful information on the interrelations among the interested set of variables. Since the comparisons among proficiency dimensions are not useful in testing practice, one can standardize the set of proficiency components and thus make the variance for each proficiency dimension equal to 1, which reduce the number of parameters in the proficiency covariance. For example, if a test requires 3 dimensional proficiency, three additional parameters are needed to describe the proficiency covariance. However, the off-diagonal components represent the interrelations among the required proficiency dimensions and the pair-wise correlations in the matrix may vary. For the multi-dimensional MIRT model, the generating covariance matrices used are in the form of

$$\begin{pmatrix} 1 & \rho & \cdots & \rho \\ \vdots & 1 & \cdots & \rho \\ \rho & \cdots & \ddots & \rho \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

for simplicity, where  $\rho$  in the proficiency structure matrix equals to .2, which is denoted as,

$$\Sigma_{\theta,2} \equiv \begin{pmatrix} 1 & .2 & \cdots & .2 \\ \vdots & 1 & \cdots & .2 \\ .2 & \cdots & \ddots & .2 \\ .2 & .2 & \cdots & 1 \end{pmatrix}.$$

For a more general case,  $\rho$  takes different values for the off-diagonal components. For example, the generating covariance matrix for the 3-dimensional model has off-diagonal components from .2 to .7 denoted as

$$\Sigma_{\theta,g} \equiv \begin{pmatrix} 1 & .7 & .2 \\ .7 & 1 & .3 \\ .2 & .3 & 1 \end{pmatrix}.$$

### 3.5.4 Generating Item Parameters

It is natural to assume that some items in a test only measure one dimension proficiency (call such items uni-items), some items may measure two or more dimensions (call such items multi-items). A test can be composed by both uni-items and multi-items. Two tests that include both uni-items and multi-items are generated in this simulation study on estimation for the 3-dimensional MIRT model with 30 and 45 items, respectively.

Table 3.10 contains the true item parameters for the 30-item test. The first 15 items only measure one dimension proficiency and the remaining 15 items measure three dimension abilities. The parameter vector  $\mathbf{a}$  ranges from 0 to 2.45. Note for the items which measure 3-dimensional abilities, some components in the  $\mathbf{a}$  parameter are dominant over other dimensions (e.g., item 20, 21, 24), and some items have very close values of  $\mathbf{a}$  parameters on two or three dimensions (e.g., item 19, 25, 26, 27). The values of  $\mathbf{d}$  parameters are simulated from the standard normal distribution  $N(0, 1)$ .

The lowest  $d$  value is -1.63 and the highest value of  $d$  is 2.38, indicating a wide range of  $d$  values is included in the test. Asymptote parameters  $c$  are drawn from the uniform distribution  $U(0, .25)$ . High guessing parameters are not expected for good test items, as in the case of this example. Combined with the number of items (e.g., 30 and 45 items) and the sample size (e.g., 2000 and 5000), and the underlying proficiency structure (e.g.,  $\Sigma_{\theta,2}$  and  $\Sigma_{\theta,g}$ ), there are in all 24 dichotomous response data sets generated.

To solve the indeterminacy problem and establish a fixed scale for the model parameter estimates, the first three items are chosen as an unidimensional item, which is strongly considered to measure only the first, the second, and the third dimension, respectively. More specifically, the  $a$  values for the first item takes zero on the second and third dimensions, the  $a$  values for the second item takes zero on the first and third dimensions, and similarly the  $a$  values for the third item takes zero on the first and second dimensions. These three items are viewed as *anchor* items, because they are placed at the first three positions in the test and all are uni-dimensional items, which is treated as a constraint in order to settle the metric issue or the indeterminacy problems that are inherited in the MIRT models. It is argued that the model can be identified by setting the mean vector of proficiency parameters equal to zero and standardizing the covariance matrix, plus the above constraints, which are also used in the exploratory option of NOHARM (Fraser, 1988).

Table 3.11 contains the true parameters for the 45-item test. The first thirty items only measure one dimension proficiency. Item 1 and item 4 to item 12 only

load on the first dimension, item 2 and item 13 through item 21 measure the second dimension, and item 3 and item 22 through item 30 only load on the third dimension. The remaining 15 items of the test, item 31 through item 45, are able to measure all three dimensions. The parameters  $a$  in the 45-item test also see a wide range as well, from 0 (item 2) to 2.43 (item 41). The minimum value of parameter  $d$  is -2.06 (item 26) and the maximum is 2.07 (item 6). The parameters  $c$  are within the range of .01 to .24 in this test.

The first three items in the 45-item test are also uni-dimensional items and placed in the first three positions in the test, which is to believe that these three items are able to measure well the first, the second, and the third dimension, respectively. The purpose of placing the three uni-dimensional item in the first three positions in the test is to settle the indeterminacy problems and establish a fixed scale for the item and proficiency parameter estimates.

### **3.5.5 The Estimation Accuracy and Stability for the 3-Dimensional MIRT Model**

Table 3.12 contains the RMSE for the item parameters in the 3-dimensional model for both tests with the sample size 2000 and 5000 and in a condition that all of the off-diagonal components for the proficiency covariance are equal to .2. Note the item parameter estimates are the means of the three individual estimates of the item parameters, which are based on the three chains with different random initial values. By taking the means of the individual estimates based on multiple chains for the same data set, one can expect the the final estimates to be more stable and accurate

Table 3.10: True Item Parameters for 30-Item Test (Dim = 3)

Item	$a_1$	$a_2$	$a_3$	$d$	$c$
1	1.30	0	0	-0.23	.21
2	0	0.50	0	0.02	.00
3	0	0	2.10	-1.00	.24
4	1.93	0	0	0.61	.06
5	0.81	0	0	0.31	.24
6	1.62	0	0	1.76	.00
7	0.59	0	0	1.56	.06
8	0	2.45	0	-0.38	.08
9	0	1.88	0	-0.86	.14
10	0	0.57	0	-0.51	.02
11	0	1.15	0	1.25	.03
12	0	0	1.35	-0.29	.25
13	0	0	0.98	2.38	.09
14	0	0	1.46	-1.45	.12
15	0	0	1.49	-0.30	.21
16	1.34	2.23	1.98	-1.24	.05
17	1.84	2.34	0.90	0.08	.00
18	0.86	1.04	1.76	1.13	.03
19	1.93	1.65	1.96	0.61	.11
20	0.56	0.87	1.97	1.23	.09
21	2.20	0.96	1.16	-1.01	.05
22	1.58	1.48	2.29	-1.58	.19
23	1.26	1.68	1.45	-0.07	.16
24	2.37	0.75	0.52	-1.37	.02
25	1.94	1.99	1.16	-1.63	.04
26	0.89	1.32	0.92	0.35	.20
27	1.25	1.56	1.64	1.08	.06
28	2.07	1.71	2.43	0.79	.06
29	1.41	0.96	2.12	0.46	.20
30	0.98	2.30	1.64	-0.43	.08

Table 3.11: True Item Parameters for 45-Item Test (Dim = 3)

Item	$a_1$	$a_2$	$a_3$	$d$	$c$	Item	$a_1$	$a_2$	$a_3$	$d$	$c$
1	1.12	0	0	0.18	.09	24	0	0	0.24	0.56	.13
2	0	1.51	0	1.28	.08	25	0	0	1.96	-0.23	.14
3	0	0	1.24	-0.46	.19	26	0	0	0.44	-2.06	.04
4	2.03	0	0	-1.74	.05	27	0	0	0.91	0.24	.07
5	1.92	0	0	1.24	.14	28	0	0	2.46	0.05	.07
6	1.84	0	0	2.07	.19	29	0	0	0.77	1.06	.01
7	2.43	0	0	0.42	.11	30	0	0	1.57	0.51	.02
8	0.94	0	0	1.04	.03	31	2.26	0.52	1.85	-2.05	.06
9	0.89	0	0	0.27	.13	32	1.20	1.05	0.79	0.28	.04
10	0.52	0	0	-0.69	.22	33	2.31	1.25	1.98	-0.31	.13
11	0.30	0	0	-0.75	.23	34	1.38	0.64	1.62	0.80	.02
12	0.94	0	0	0.65	.12	35	0.53	2.21	1.23	-0.55	.02
13	0	0.53	0	-0.92	.22	36	0.95	1.09	1.02	-0.99	.23
14	0	0.91	0	1.28	.03	37	0.22	0.92	0.80	0.04	.23
15	0	0.22	0	0.02	.17	38	1.77	2.50	0.78	1.33	.01
16	0	1.03	0	-1.64	.02	39	1.32	2.19	1.32	1.30	.13
17	0	1.87	0	-1.69	.02	40	1.85	1.08	1.22	-0.45	.09
18	0	0.79	0	-1.11	.03	41	1.27	2.21	2.43	1.98	.10
19	0	1.81	0	-0.47	.17	42	0.22	1.15	2.00	0.50	.24
20	0	1.75	0	1.31	.12	43	0.36	1.25	0.21	-0.43	.16
21	0	0.73	0	-1.07	.21	44	1.95	1.60	1.35	-0.90	.03
22	0	0	2.05	-1.22	.04	45	1.11	2.21	1.07	-0.40	.07
23	0	0	2.05	0.47	.05	-	-	-	-	-	-

Table 3.12: RMSE for Multi-dimensional Test (Dim = 3,  $\rho = .2$ )

Estimates	$30 \times 2000$	$30 \times 5000$	$45 \times 2000$	$45 \times 5000$
$\hat{a}_1$	.15	.08	.15	.06
$\hat{a}_2$	.12	.08	.16	.04
$\hat{a}_3$	.18	.08	.11	.05
$\hat{d}$	.19	.10	.22	.11
$\hat{c}$	.07	.03	.06	.03



because the fluctuation of the item parameter estimates induced by the initial values and sampling errors are taken into account. It shows for a given test (e.g., the 30-item or 45-item test) the larger the sample size, the smaller RMSE. For the 30-item test, the largest RMSE for  $a$  is .18 when sample size is 2000, but is .8 when sample size is 5000. The RMSE for  $d$  parameter is .19 when sample size is 2000, and is .10 for the sample size 5000. The RMSE for  $c$  parameter is .07 for sample size 2000, but is .03 for 5000. Similar results can also be found in the 45-item test. The smallest RMSE is .04 for  $a$  parameter in the 45-item test with sample size 5000. Note that within the same test and with the same sample size, the RMSE for  $a_i, \forall i = 1, 2, 3$  are close to each other, which implies that the estimation can achieve the same level of precision across dimensions. It also shows that the RMSE for  $c$  is generally smaller than the RMSE for  $a$  and  $b$ , with the largest one .07 in the 30-item test to 2000 examinees.

Table 3.13 gives the RMSE for the situation in which the underlying proficiency covariance is a general one or it does not follow a special pattern (e.g., all off-diagonal components on the proficiency covariance matrix are the same). The results of the parameter estimation for this particular condition are found very similar to the case in which the off-diagonal components for the covariance matrix are equal to .2. This implies that the underlying proficiency covariance does not affect the item parameter estimates, which is expected because the estimation of the item and proficiency parameters are independent.

Compared to the RMSE for the unidimensional model in Table 3.8, the RMSE

Table 3.13: RMSE for Multi-dimensional Test (Dim =3,  $\rho$  = general)

Estimates	30 × 2000	30 × 5000	45 × 2000	45 × 5000
$\hat{a}_1$	.10	.12	.13	.06
$\hat{a}_2$	.14	.06	.11	.07
$\hat{a}_3$	.13	.12	.13	.06
$\hat{d}$	.13	.11	.21	.15
$\hat{c}$	.06	.03	.06	.04

Table 3.14: Correlations Between True Proficiency and Estimates (Dim = 3,  $\rho$  = .2)

	30 × 2000	30 × 5000	45 × 2000	45 × 5000
$\text{corr}(\theta_1, \hat{\theta}_1)$	.8765	.8737	.9144	.9136
$\text{corr}(\theta_2, \hat{\theta}_2)$	.8677	.8703	.9125	.9121
$\text{corr}(\theta_3, \hat{\theta}_3)$	.8531	.8649	.9109	.9146

for item parameter estimates in Table 3.12 and 3.13 are generally higher those item parameter estimates for the 3-dimensional MIRT model. It is clear that given the same size of data information, the more parameters to be estimated, the more estimation errors.

It can be seen that for the same test, larger sample size gives smaller RMSE. The RMSE for  $a$  parameter cross dimensions are close to each other with a range from .10 to .14 for the sample size 2000 and a range of .06 to .12 for the sample size 5000. The largest RMSE for  $d$  is .21, which occurs in the 45-item test with 2000 examinees, the smallest is .11 in the 30-item test with sample size 5000. Generally speaking, The RMSE for parameter  $c$  are smaller than those for parameters  $a$  and  $d$ , varying from .03 to .06, because  $c$  is restricted to a very small range. The RMSE of  $c$  for the sample size 5000 are about the half of the ones for 2000 examinees.

Table 3.15: Correlations Between True Proficiency and Estimates (Dim = 3,  $\rho$  = general)

	30 × 2000	30 × 5000	45 × 2000	45 × 5000
$\text{corr}(\theta_1, \hat{\theta}_1)$	.8876	.8943	.9198	.9259
$\text{corr}(\theta_2, \hat{\theta}_2)$	.8878	.8966	.9211	.9255
$\text{corr}(\theta_3, \hat{\theta}_3)$	.8474	.8602	.9111	.9101

The correlations between true abilities and estimates are presented in Table 3.14 and 3.15 for  $\rho = .2$  and  $\rho$  is varied, respectively. Table 3.14 shows that for the 30-item test the correlation between the true values and the estimates are around .87 with a very small range from .8531 to .8765. Also, the correlations for the 45-item test slightly differ from .9109 to .9146. The 45-item test in general has higher correlations (around .91) between the true and the estimated abilities than those in the 30-item tests. This implies the proficiency estimates get improved for the longer test, or the estimation precision for proficiency in the longer test is better than that in the short test (i.e., the 30-item test).

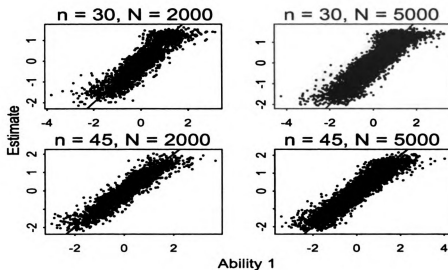
Table 3.15 presents the correlations between the true proficiency ( $\theta$ ) and the estimates ( $\hat{\theta}$ ) for the situation in which the components for the off-diagonal proficiency covariance matrix take different values. The 30-item test gives correlations from .8474 to .8966. Higher correlations are also found in the 45-item tests with a range from .9101 to .9259. No noticeable difference of correlations have been found cross dimensions. For example, for the 30-item test with 2000 examinees, the correlation between the first proficiency dimension and its estimates,  $\text{corr}(\theta_1, \hat{\theta}_1) = .8876$ , the correlation between the second proficiency dimension and its estimates,  $\text{corr}(\theta_2, \hat{\theta}_2) = .8878$ , and

the correlation for the third dimension is  $\text{corr}(\theta_3, \hat{\theta}_3) = .8474$ . Comparing Table 3.14 to 3.15, slightly higher correlations appear in the situation that  $\rho$  takes different values than the fixed  $\rho = .2$  condition. But the difference is negligible.

In general, the correlations for the unidimensional model in Table 3.9 are higher than those for the 3-dimensional model in Table 3.14 and 3.15. This implies that as the number of dimensions increases from 1 to 3, the number of parameters to be estimated increases from 2090 to 6153 for the 30-item test to 2000 examinees. Therefore, more estimation errors will appear in the item and proficiency estimates for the 3-dimensional model.

Figure 3.7 through Figure 3.9 show the plots of the true proficiency versus the estimates for the 30-item and the 45-item tests cross different sample sizes. The plots in these 3 figures demonstrate that the true and estimates are more close to the reference line  $y = x$  for the longer test (45-item), as is consistent with the findings on the correlations in Table 3.14 and 3.15. Figure 3.10 through 3.13 are the plots of the true item parameters versus their estimates and they are all tightly around the reference line, showing the stable and accurate estimates are obtained in various simulation conditions regarding the test length, the examinee sample size, and the underlying proficiency covariance. It is worth pointing out that from the Figure 3.10, 3.11, and 3.12, for  $a$  parameters with true value 0, the estimates are close to zero. The estimates in the tests with larger sample size (*e.g.*,  $N = 5000$ ) are even closer to zero, with the biggest difference between the true parameters and estimates less than .2.

Figure 3.7: True Proficiency Versus Estimates ( $Dim = 3, \rho = general, n = 30, N = 5000$ )



Note that the plots are for the situation in which the underlying proficiency covariance matrix is  $\Sigma_{\theta,2}$ . Similar results are also obtained when the proficiency covariance is  $\Sigma_{\theta,g}$ , in which the pairwise correlations vary, but the plots are omitted here.

### 3.6 Estimating the 5-dimensional Model

The two tests for the simulation studies in this section will have the same number of item (e.g.,  $n = 30$  or  $n = 45$ ) and will also be administrated to the groups of examinees with size  $N = 2000$  and  $N = 5000$ , respectively. The differences are both tests are assumed to require five dimensions of proficiency to correctly answer the items in the two tests. Since the tests are to measure five dimensions of abilities, the total number

Figure 3.8: True Proficiency Versus Estimates ( $Dim = 3, \rho = general, n = 45, N = 2000$ )

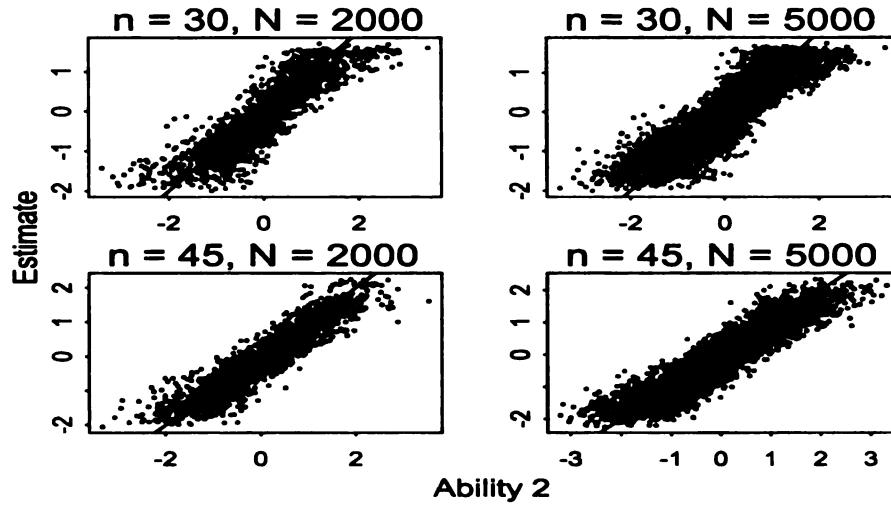


Figure 3.9: True Proficiency Versus Estimates ( $Dim = 3, \rho = general, n = 45, N = 2000$ )

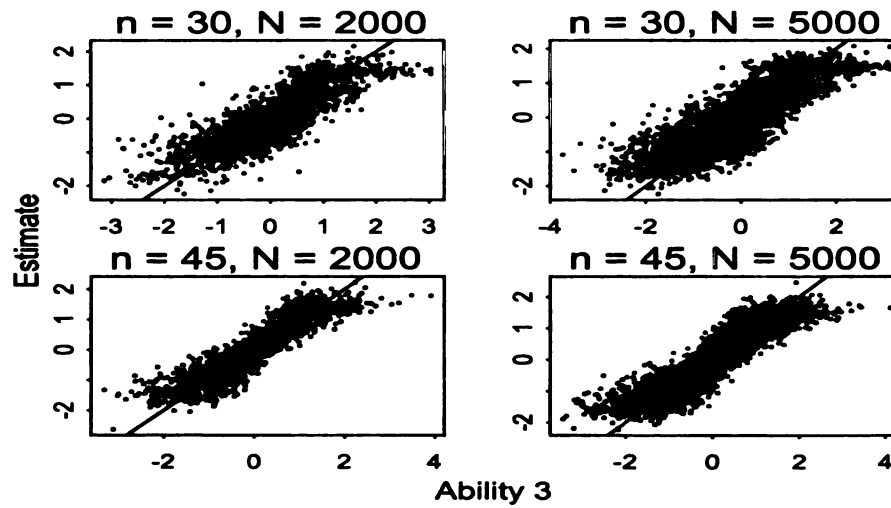


Figure 3.10: True  $a_1$  Parameter Versus Estimates ( $Dim = 3, \rho = .2$ )

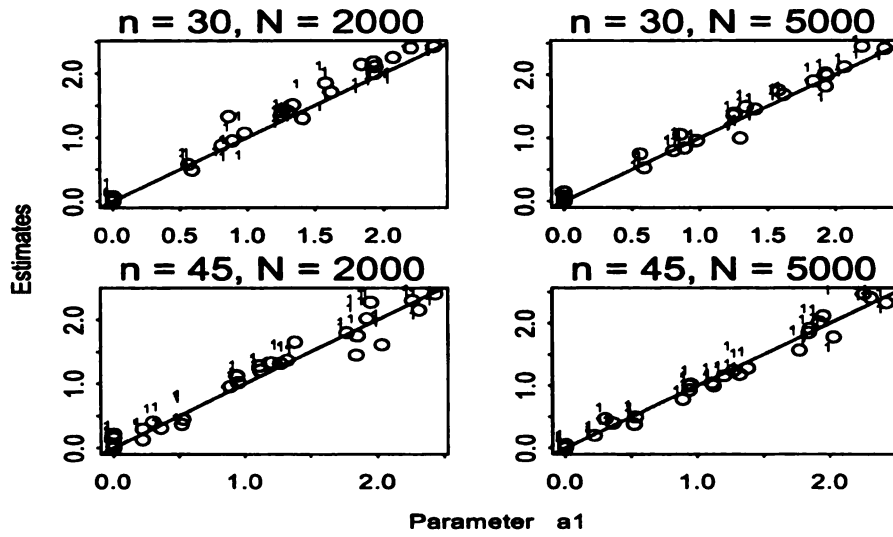


Figure 3.11: True  $a_2$  Parameter Versus Estimates ( $Dim = 3, \rho = .2$ )

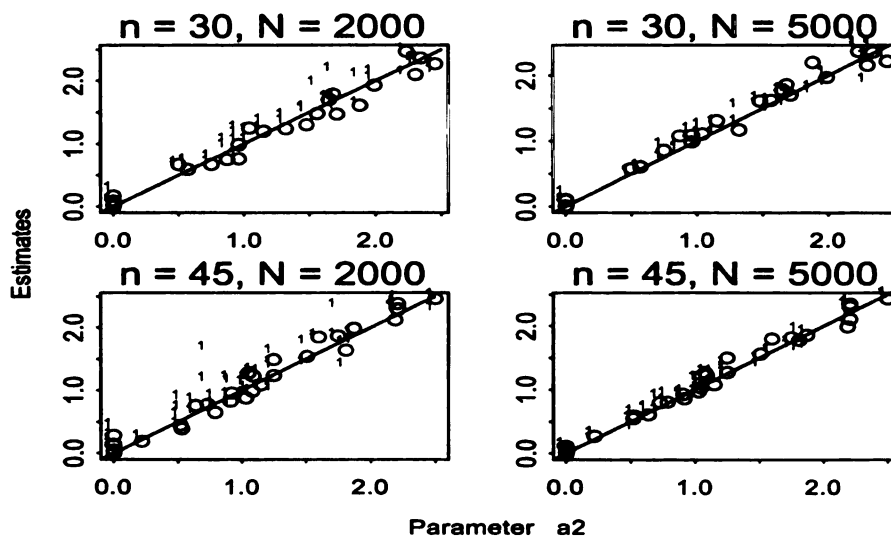


Figure 3.12: True  $a_3$  Parameter Versus Estimates ( $Dim = 3, \rho = .2$ )

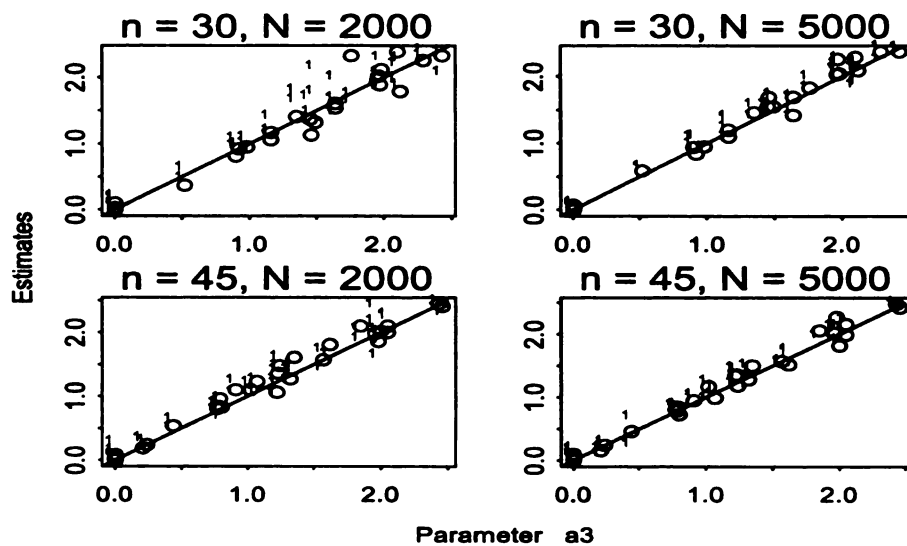
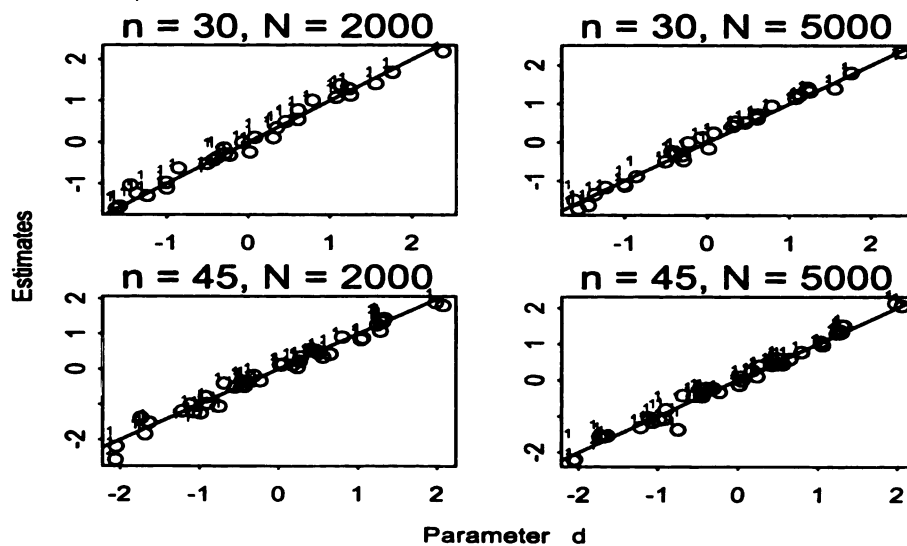


Figure 3.13: True  $d$  Parameter Versus Estimates ( $Dim = 3, \rho = .2$ )





of parameters to be estimated are  $(5 + 2)n + 5N + 10$ , where  $n$  stands for test length and  $N$  for the sample size of examinee. For the 30-item test that is administrated to 2000 examinees, for example, the total number of model parameters need to be estimated from the observed data is 10220, which is much greater than the sample size 2000. If this test is to administrated to a group of 5000 examinees, the number of model parameters is 25220. Similarly, for a 45-item test that is administrated to a group of 2000 examinees, the total number of model parameters is 10325, and is 25325 if administrated to a sample of 5000 examinees.

The design for the 30-item test that is assumed to measure five dimensions of abilities will follow the same pattern as that of the three dimensional tests. To put some constraints for the model identification and the establishment of the fixed scale for the parameter estimates, the first five items are unidiemsional items and are placed on the first five positions in the test with each item measuring only one dimension of proficiency. More specifically, these items are also called anchor items with the first item only measuring the first dimension of proficiency and the second items only measuring the second dimension, and so on. Table 3.16 and 3.17 contain the true item parameters for the 30-item test and the 45-item test, respectively. It can be seen that the anchor items have a wide range of values on the  $a$  parameters (e.g., from .65 to 2.04 for the 30-item test, and from 1.38 to 2.32 for the 45-item test). In the 30-item test, there are two additional unidimensional items (e.g., item 6 through item 15) for each dimension and the rest of the items are assumed to measure all five dimensions of abilities (e.g., item 16 through item 30). For the 45-item test,

only one additional unidimensional item for each dimension are present in the test, item 6 through item 10. The rest of the items in this test are suppose to measure all five dimensions of abilities. In the 30-item test, each dimension of proficiency is designed to be measured by only 17 items. And in the 45-item test, each dimension of proficiency can be measured by 42 items, much more than that in the 30-item test. According to this design of items for the two tests, one would reasonable expect that the proficiency estimates in the 45-item would be improved since more items are designed to measure each dimension of proficiency.

Note that the true item parameters in both tests in Table 3.16, 3.17 and 3.18 include a wide range of values on each item parameter. For example, the largest value of  $a$  parameter is 2.44 and the lowest is 0 in the 30-item test, and the largest and lowest  $a$  values in the 45-item test are 2.32 and 0, respectively. The values on  $d$  parameters for both tests have a reasonable range, which are both from a standard normal distribution. All the asymptote parameters are controlled within the range between 0 and .3.

The five dimensional proficiency parameters are randomly generated from a multivariate normal distribution with the mean vector  $\mathbf{0}$  and the covariance matrix  $\Sigma_{\theta}$  (i.e.,  $N(\mathbf{0}, \Sigma_{\theta})$ ). As in the case for the three dimensional tests in Section 3.5, the mean vector for the underlying proficiency distribution is set to  $\mathbf{0}$  to establish the same scale for each proficiency dimension. In the same way, the covariance matrix  $\Sigma_{\theta}$  is standardized and becomes actually the correlation matrix among these dimensions of abilities. The pairwise correlation among these five dimensions (or the off-diagonal

Table 3.16: True Item Parameters for 30-Item Test (Dim = 5)

Item	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$d$	$c$
1	0.65	0	0	0	0	1.76	.20
2	0	1.74	0	0	0	-0.69	.23
3	0	0	2.04	0	0	0.13	.15
4	0	0	0	1.38	0	1.13	.24
5	0	0	0	0	0.98	-0.64	.14
6	1.14	0	0	0	0	0.30	.07
7	1.64	0	0	0	0	-0.11	.10
8	0	0.67	0	0	0	-0.62	.23
9	0	1.21	0	0	0	0.73	.25
10	0	0	1.49	0	0	-1.12	.12
11	0	0	0.99	0	0	-1.10	.12
12	0	0	0	1.18	0	1.34	.04
13	0	0	0	1.41	0	2.02	.09
14	0	0	0	0	1.91	0.49	.16
15	0	0	0	0	0.88	-1.28	.24
16	2.44	1.24	2.18	1.88	0.85	0.85	.03
17	1.81	1.85	2.28	1.21	2.44	-1.64	.13
18	1.02	2.14	1.77	1.80	2.02	0.91	.07
19	0.60	1.75	2.14	2.19	2.35	2.73	.03
20	0.94	1.23	2.07	1.91	1.42	1.43	.19
21	1.01	1.39	2.17	2.26	0.98	0.95	.12
22	1.13	1.47	2.50	1.08	1.84	2.30	.08
23	1.32	1.29	1.59	2.20	0.80	0.48	.22
24	0.73	2.28	2.00	0.86	0.87	0.51	.15
25	2.43	1.08	1.84	1.15	2.03	0.20	.15
26	1.73	1.30	2.42	1.29	1.15	0.21	.00
27	1.98	1.69	1.50	2.28	1.46	-0.71	.15
28	2.00	1.39	2.15	0.59	1.10	-0.86	.09
29	1.62	1.92	1.56	2.07	1.91	-0.09	.10
30	0.81	1.70	2.13	1.39	1.28	0.75	.06

Table 3.17: True Item Parameters for 45-Item Test (Dim = 5)

Item	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$d$	$c$
1	2.32	0	0	0	0	-0.17	.18
2	0	1.94	0	0	0	0.16	.22
3	0	0	1.53	0	0	0.36	.13
4	0	0	0	1.38	0	0.30	.19
5	0	0	0	0	1.71	0.47	.12
6	1.51	0	0	0	0	0.71	.23
7	0	1.74	0	0	0	-1.61	.21
8	0	0	1.90	0	0	-0.88	.03
9	0	0	0	2.14	0	-1.15	.16
10	0	0	0	0	1.34	-0.13	.18
11	1.30	1.61	1.93	2.05	0.83	-0.73	.00
12	1.03	2.24	0.73	2.20	1.94	2.12	.23
13	2.05	1.56	1.09	0.92	1.83	-0.75	.04
14	1.36	0.93	0.90	1.89	1.45	1.12	.17
15	1.42	2.11	0.88	1.22	0.80	-0.07	.20
16	1.10	0.95	1.83	0.80	1.34	0.00	.23
17	1.65	1.52	2.15	1.09	1.38	1.01	.15
18	1.48	1.25	1.00	1.19	1.85	2.17	.14
19	0.82	1.49	0.62	2.01	1.84	-0.58	.21
20	0.87	1.79	1.61	1.10	1.31	-0.92	.02
21	0.93	2.07	1.49	1.11	1.85	0.80	.03
22	1.03	2.14	1.76	2.33	1.49	0.01	.01

Table 3.18: True Item Parameters for 45-Item Test (Dim = 5), cont.

Item	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$d$	$c$
23	1.97	2.17	2.32	2.10	1.57	-0.44	.08
24	1.79	1.25	1.93	1.87	2.34	0.17	.24
25	1.29	0.76	2.20	1.70	1.60	-1.35	.10
26	1.50	1.90	2.03	1.31	1.07	-0.74	.17
27	1.65	0.90	1.42	1.81	0.69	-0.31	.12
28	2.31	0.82	1.91	1.50	1.75	-2.08	.19
29	0.93	2.35	2.34	1.70	1.12	0.36	.08
30	1.99	0.73	1.58	1.68	1.04	-1.36	.08
31	1.34	1.20	1.88	2.18	1.60	-0.81	.18
32	1.49	1.50	1.76	2.00	1.63	-0.25	.12
33	1.95	2.22	1.39	1.59	1.09	-0.29	.11
34	0.64	1.26	0.80	1.21	0.95	-1.55	.23
35	1.06	1.51	1.69	1.64	1.17	-0.60	.09
36	1.45	0.82	1.92	1.66	0.49	0.50	.13
37	1.52	2.22	0.87	1.70	0.71	0.82	.13
38	2.04	1.45	0.97	2.28	1.81	0.96	.24
39	0.90	2.06	1.27	1.55	1.25	1.83	.00
40	1.93	2.09	1.65	1.25	0.80	0.78	.04
41	1.44	1.01	0.81	2.13	1.22	0.19	.25
42	0.74	1.78	1.94	0.92	2.07	-1.01	.04
43	1.93	1.81	0.69	0.90	1.79	0.08	.09
44	1.11	1.91	1.83	0.86	1.06	1.60	.22
45	2.05	1.25	1.55	0.89	1.79	0.89	.02

components in  $\Sigma_{\theta}$ ) can be the same or can vary from each other. In this section, two covariance matrices of  $\Sigma_{\theta}$  are used and denoted as  $\Sigma_{\theta,2}$  and  $\Sigma_{\theta,g}$ , respectively. From the notations on the covariance matrices, one can see that the former covariance matrix indicates that all the off-diagonal components take the same values (e.g., .2) and the off-diagonal components for the latter covariance matrix vary from .2 to .6, which is shown as

$$\Sigma_{\theta,g} \equiv \begin{pmatrix} 1 & .2 & .3 & .2 & .5 \\ .2 & 1 & .2 & .5 & .3 \\ .3 & .2 & 1 & .4 & .3 \\ .2 & .5 & .4 & 1 & .6 \\ .5 & .3 & .3 & .6 & 1 \end{pmatrix}.$$

Combined with the test length (30 and 45), the sample size (2000 and 5000), the proficiency covariance ( $\Sigma_{\theta,2}$  and  $\Sigma_{\theta,g}$ ), and the replications, 24 response data sets are yielded for the simulation studies on the five dimensional case. For each data set, multiple chains (e.g., 3 chains for each data set) will be constructed. To give more stable and accurate estimates, the final estimates for item parameters will take the means of the three individual estimates from each chain with different initial values. Therefore, there are in all 72 runs for the parameter estimates in this section.

Table 3.19 and Table 3.20 give the RMSE for the item parameter estimates for the eight simulation conditions for each item parameter. The differences between the two tables are that the underlying proficiency covariance is different. The results of Table 3.19 are based on  $\Sigma_{\theta,2}$  and Table 3.20 on  $\Sigma_{\theta,g}$ . Most of the RMSE in the tables are less than .2. The highest RMSE value (.29) is for  $d$  parameter in the condition of 5000 examinee on the 45-item test with covariance  $\Sigma_{\theta,g}$ .

Clearly from the two tables, the precision of item parameter estimates does not change due to the use of different proficiency covariance. Or the underlying proficiency covariance is not a factor that can affect the item parameter estimates, which is expected because sampling of item and proficiency parameters are independent. It is also clear that the RMSE are generally smaller when the sample size is 5000 than those when the sample size is 2000, which is also expected since more examinees provide more information on item parameter estimation. However, the estimation seems better on the 30-item test since the RMSE have slightly higher values in the 45-item test in general no matter what the sample size is, which is not expected. One possible reason is that the dimension structure in the 30-item test (only 17 items measuring all 5 dimensions) is much simpler than the 45-item test (32 items measuring all 5 dimensions). In addition, more items with extreme values that are difficult to estimate, might appear in the 45-item tests.

Compared to the RMSE for item parameter estimates in the unidimensional model (Table 3.8) and the 3-dimensional model (Table 3.12 and 3.13), the RMSE for the item parameter estimates for the 5-dimensional model (Table 3.19 and 3.20) are generally higher. Again, this implies for the same size of data information, the more parameters to be estimated as the number of dimensions increases, the more errors for the estimation.

Table 3.21 shows the correlations between the true and estimates of proficiency parameters when the underlying covariance matrix is  $\Sigma_{\theta,2}$ . That is, the off diagonal components for the covariance matrix of the proficiency distribution is equal to .2.

Table 3.19: RMSE for Multi-dimensional Test (Dim = 5,  $\rho = .2$ )

Estimates	$30 \times 2000$	$30 \times 5000$	$45 \times 2000$	$45 \times 5000$
$\hat{a}_1$	.15	.13	.21	.20
$\hat{a}_2$	.24	.16	.20	.14
$\hat{a}_3$	.16	.11	.23	.15
$\hat{a}_4$	.20	.14	.22	.22
$\hat{a}_5$	.18	.15	.20	.14
$\hat{d}$	.21	.16	.22	.24
$\hat{c}$	.05	.05	.03	.03

Table 3.20: RMSE for Multi-dimensional Test (Dim = 5,  $\rho = \text{general}$ )

Estimates	$30 \times 2000$	$30 \times 5000$	$45 \times 2000$	$45 \times 5000$
$\hat{a}_1$	.17	.15	.26	.17
$\hat{a}_2$	.18	.16	.27	.18
$\hat{a}_3$	.16	.15	.20	.19
$\hat{a}_4$	.18	.16	.25	.21
$\hat{a}_5$	.21	.18	.24	.25
$\hat{d}$	.25	.17	.28	.29
$\hat{c}$	.06	.04	.03	.03



Table 3.21: Correlations Between True Proficiency and Estimates (Dim = 5,  $\rho = .2$ )

	$30 \times 2000$	$30 \times 5000$	$45 \times 2000$	$45 \times 5000$
$\text{corr}(\theta_1, \hat{\theta}_1)$	.7899	.7829	.7935	.7976
$\text{corr}(\theta_2, \hat{\theta}_2)$	.7508	.7499	.7984	.8006
$\text{corr}(\theta_3, \hat{\theta}_3)$	.8038	.8067	.8088	.8195
$\text{corr}(\theta_4, \hat{\theta}_4)$	.7606	.7641	.7934	.7818
$\text{corr}(\theta_5, \hat{\theta}_5)$	.7594	.7559	.8010	.7915

In general, the correlation for each dimension in this study is around .8, and the correlations are close between the two proficiency covariance conditions, indicating the proficiency covariance does not affect proficiency estimates. When compared to the correlations for the unidimensional model (Table 3.9) and the 3-dimensional model (Table 3.14 and 3.15), the correlations for the 5-dimensional model in Table 3.21 and 3.22 are generally smaller, which is expected because as the dimension increases, more parameters are to be estimated. The lowest correlations are for the short test (the 30-item test), which is expected, because each dimension of proficiency is measured by only 17 items. The longer test (the 45-item test) has slightly higher correlation coefficients. Low correlations indicate large estimation errors for the proficiency estimates. Nevertheless, the estimation is not significantly improved in the 45-item test although each dimension is measured by 32 items. One possible interpretation is that the parameters to be estimated substantially increase as the number of dimensions increases to five.

Table 3.22: Correlations Between True Proficiency and Estimates (Dim = 5,  $\rho$  = general)

	$30 \times 2000$	$30 \times 5000$	$45 \times 2000$	$45 \times 5000$
$\text{corr}(\theta_1, \hat{\theta}_1)$	.7835	.7935	.8076	.7999
$\text{corr}(\theta_2, \hat{\theta}_2)$	.7548	.7617	.7882	.7983
$\text{corr}(\theta_3, \hat{\theta}_3)$	.8098	.8034	.8221	.8139
$\text{corr}(\theta_4, \hat{\theta}_4)$	.8171	.8241	.8264	.8326
$\text{corr}(\theta_5, \hat{\theta}_5)$	.7745	.7971	.8244	.8333

### 3.7 Proficiency Structure Estimation

The estimates of the underlying proficiency structure have potential affects on the convergence speed, since at each sampling step, the proficiency samples are taken from the multivariate normal distribution with mean vector  $\mathbf{0}$  and sample covariance from the inverse Whishart distribution based on the sample covariance of abilities. Good recovery of the covariance structure can make an effective Markov chain.

The components of the underlying proficiency covariance are also estimated along with item and proficiency parameters by the MCMC procedure. For each data set, one estimate of covariance can be obtained for each chain replication with different initial values. The final covariance matrix estimate is the mean of the three estimates from independent chains. Note for each chain, the proficiency covariance estimate is the mean of the 1000 sample of the covariance from inverse Wishart distribution, which is also based on the sample covariance. The good estimates of covariance matrix would better recover the interrelations across proficiency dimensions. Table 3.23 gives estimates for each chain of the 30-item test in three-dimensional case with

Table 3.23: Estimates of Covariance Matrix, Dim = 3,  $\rho = .2$ 

Data	30 × 2000	30 × 5000
Rep 1	$\begin{pmatrix} 1.02 & 0.21 & 0.15 \\ & 1.01 & 0.14 \\ & & 0.97 \end{pmatrix}$	$\begin{pmatrix} 1.01 & 0.15 & 0.13 \\ & 1.03 & 0.13 \\ & & 0.98 \end{pmatrix}$
Rep 2	$\begin{pmatrix} 1.04 & 0.18 & 0.17 \\ & 0.99 & 0.12 \\ & & 0.96 \end{pmatrix}$	$\begin{pmatrix} 0.99 & 0.18 & 0.18 \\ & 1.00 & 0.18 \\ & & 1.01 \end{pmatrix}$
Rep 3	$\begin{pmatrix} 0.99 & 0.13 & 0.17 \\ & 1.03 & 0.21 \\ & & 1.01 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.12 & 0.14 \\ & 1.01 & 0.17 \\ & & 1.00 \end{pmatrix}$

Table 3.24: Estimates of Covariance Matrix, Dim = 3,  $\rho = \text{general}$ 

Data	45 × 2000	45 × 5000
Rep 1	$\begin{pmatrix} .95 & .58 & .15 \\ & .94 & .29 \\ & & .98 \end{pmatrix}$	$\begin{pmatrix} .99 & .69 & .16 \\ & 1.00 & .25 \\ & & 1.01 \end{pmatrix}$
Rep 2	$\begin{pmatrix} 1.04 & .65 & .13 \\ & 1.02 & .24 \\ & & .99 \end{pmatrix}$	$\begin{pmatrix} 1.01 & .68 & .14 \\ & 1.01 & .25 \\ & & 1.01 \end{pmatrix}$
Rep 3	$\begin{pmatrix} .99 & .60 & .18 \\ & .97 & .22 \\ & & 1.05 \end{pmatrix}$	$\begin{pmatrix} .98 & .70 & .19 \\ & 1.02 & .27 \\ & & .99 \end{pmatrix}$

$\rho$  taking the same value of .2 for all off-diagonal components. The table shows the diagonal elements are all close to 1, ranging from .96 to 1.04. The off diagonal elements ranges from .12 to .21. Similarly, Table 3.24 shows the covariance estimate for the 45-item test in the three-dimensional case with true covariance  $\Sigma_{\theta, g}$ . Clearly, the estimate of each component is close to their true parameter. Results from the five dimensional case in Table 3.25 and 3.26 also indicate the reasonably good recovery of the proficiency structure.

Table 3.25: Estimates of Covariance Matrix, Dim = 5,  $\rho$  = general

Data	30 × 2000	30 × 5000
Rep 1	$\begin{pmatrix} 1.03 & .14 & .24 & .23 & .47 \\ & 1.05 & .18 & .39 & .23 \\ & & 1.00 & .29 & .19 \\ & & & .99 & .54 \\ & & & & 1.05 \end{pmatrix}$	$\begin{pmatrix} 1.00 & .26 & .26 & .21 & .47 \\ & .99 & .22 & .46 & .25 \\ & & 1.00 & .32 & .31 \\ & & & .99 & .43 \\ & & & & .97 \end{pmatrix}$
Rep 2	$\begin{pmatrix} 1.01 & .27 & .30 & .28 & .44 \\ & 1.03 & .28 & .55 & .29 \\ & & 1.02 & .39 & .25 \\ & & & 1.07 & .44 \\ & & & & 1.02 \end{pmatrix}$	$\begin{pmatrix} 1.01 & .20 & .29 & .26 & .39 \\ & 1.03 & .07 & .54 & .20 \\ & & .98 & .35 & .24 \\ & & & 1.03 & .52 \\ & & & & 1.00 \end{pmatrix}$
Rep 3	$\begin{pmatrix} 1.03 & .10 & .17 & .22 & .55 \\ & 1.01 & .25 & .45 & .21 \\ & & 1.03 & .37 & .30 \\ & & & 1.00 & .50 \\ & & & & 1.03 \end{pmatrix}$	$\begin{pmatrix} 1.01 & .20 & .30 & .22 & .53 \\ & 1.03 & .21 & .44 & .20 \\ & & .98 & .45 & .29 \\ & & & 1.03 & .41 \\ & & & & 1.00 \end{pmatrix}$

Table 3.26: Estimates of Covariance Matrix, Dim = 5,  $\rho$  = .2

Data	45 × 2000	45 × 5000
Rep 1	$\begin{pmatrix} .99 & .18 & .23 & .38 & .21 \\ & 1.05 & .22 & .29 & .06 \\ & & 1.00 & .31 & .21 \\ & & & .98 & .25 \\ & & & & 1.05 \end{pmatrix}$	$\begin{pmatrix} 1.02 & .21 & .20 & .33 & .18 \\ & .99 & .21 & .24 & .22 \\ & & 1.01 & .28 & .15 \\ & & & .99 & .14 \\ & & & & .97 \end{pmatrix}$
Rep 2	$\begin{pmatrix} 1.04 & .17 & .21 & .17 & .21 \\ & 1.02 & .15 & .28 & .21 \\ & & 1.04 & .30 & .18 \\ & & & 1.00 & .27 \\ & & & & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.02 & .16 & .18 & .33 & .15 \\ & .99 & .24 & .27 & .18 \\ & & 1.01 & .26 & .18 \\ & & & .99 & .17 \\ & & & & .97 \end{pmatrix}$
Rep 3	$\begin{pmatrix} 1.03 & .14 & .24 & .28 & .35 \\ & .98 & .32 & .22 & .13 \\ & & 1.03 & .35 & .26 \\ & & & 1.04 & .33 \\ & & & & 1.03 \end{pmatrix}$	$\begin{pmatrix} 1.01 & .19 & .20 & .33 & .17 \\ & 1.03 & .24 & .26 & .17 \\ & & .98 & .25 & .18 \\ & & & 1.04 & .17 \\ & & & & 1.00 \end{pmatrix}$

### 3.8 Computing Time

One open criticism to the MCMC approach is the extensive computation, which may depends on the program efficiency, the size of the data, the convergence speed, and the computer equipment as well. The program efficiency includes the design and algorithm in the source codes. Many researchers now use the application softwares (e.g., WINBUG, BUGS, SAS, SPLUS, MATLAB) to run MCMC procedures (e.g., Patz and Junker use S-PLUS, 1999a; Bolt uses WinBug, 2004). Some researchers use computer languages (e.g., S, R, FORTRAN, JAVA) to code their own programs. In this study, the code is written by C++ with efficient algorithm using MCMC for computing IRT model parameter estimation. The size of data involves the test length, the sample size of examinees, and number of dimensions and parameters to be estimated. In general, the longer the test, the more time is needed. Similarly, the larger number of examinees and dimensions of proficiency required, the longer the computing time is required. For a given data set, the more parameters are to be estimated, the longer the computing time is needed. As for the convergence speed, it is associated with the priors chosen for each item and proficiency parameters, and is also associated with the data structure. If each chain is diagnosed not mixed well, or not converged to the target posterior distributions, long iteration is required, and thus longer time is required. Finally, better equipped computer system give faster computation for the same program. The computing time for 11000 iterations using the C++ program is given in the Table 3.25, and it is calculated based on a computer

Table 3.27: Computing time for 1-, 3-, and 5-Dimension data

Data	$30 \times 2000$	$30 \times 5000$	$45 \times 2000$	$45 \times 5000$
1-dimension	37 min	1 hr 17 min	42 min	1 hr 33 min
3-dimension	59 min	2 hr 30 min	1 hr 20 min	3 hr 35 min
5-dimension	1 hr 35 min	4 hr 5 min	2 hr 8 min	5 hr 17 min

with 512 MB RAM and 3300 AMD Athlon 64 processor. The shortest time, 37 minutes, is in the computation of the parameter estimation for unidimensional model with 30 items and 2000 examinees. The longest time is in the case with 45 items to 5000 examinees and with 5 dimensions of proficiency, taking 5 hours and 17 minutes to finish the 10000 iterations. The time required to computing other conditions is within the range from the shortest to the longest.

## Chapter 4

# Concluding Remarks and Future Research Directions

This research involves extensive simulation studies on parameter estimation for multidimensional IRT models in various conditions in terms of the test length, the sample size of examinees, the number of dimensions, and the underlying proficiency structures using the MCMC approach. Results on parameter estimates from these conditions are compared to investigate the influence of the potential factors on the accuracy and stability of the estimation.

This study is a extensive examination on the MCMC approach to parameter estimation in terms of the test length, the examinee sample size, the number of dimensions, the proficiency covariance, the range of item parameters, and the dimensional structure in each simulated tests. For example, the study includes both unidimensional items and multidimensional items in a test, and it has a wide variety of parameter values (not limit to certain range of values for parameters). Moreover, the study does not only focus on simple structure, but also considered the complex structure.

The MCMC approach provides a convenient and flexible framework for parameter estimation of complex IRT models, as is shown in Chapter 3 for estimating multidimensional models. The C++ program is used to estimate not only the simple IRT model (e.g., unidimensional) but also some complex models (e.g., multidimensional IRT models). The framework involves estimation of any type of parameters in the IRT models (i.e., item parameters, proficiency parameters, proficiency covariance). One can use the framework to estimate both item and proficiency parameters simultaneously. Or one can obtain the estimates of the proficiency covariance matrix to infer the interrelations among the proficiency dimensions. For some simple situations, for example, if only item parameter estimates, or only proficiency estimates, or only knowing the interrelations among proficiency dimensions is required, the program can give the required estimation procedures and ignore other parameter estimation without loss of any generality. In this case, the MCMC approach would be faster because less number of parameters are to be estimated, and thus less operation time is needed. In addition, under this framework, one can give the item parameter estimates first, then treat the item parameter estimates as true to yield the proficiency parameter estimates and proficiency covariance estimates (even by other procedures, for example, ML procedure). Or one is able to estimate all the model parameters simultaneously, as is done in this study. In addition, the framework is not restricted to short tests or lower dimensional tests. It is particularly useful for estimating higher dimensional and long tests with large number of examinees, or is useful for the contexts in which the IRT model is so complicated that other estimation approaches become infeasible.



The MCMC approach is effective and the computation is efficient. For parameter estimation in unidimensional models, half an hour is enough for a test with 30 items to 2000 examinees for 11000 iterations. One hour and half to longer tests and larger sample size, for example a 45-item test to 5000 examinees. The path plots for the posterior distribution for item parameters shown in Figure 3.2 imply that the constructed chains are well mixed even in the first 3000 iterations. If some parameters are not required for the estimation, less time is needed for the estimation and the resulted estimates are not affected by ignoring other parameter estimation. For example, the item parameters estimation will take less time if no proficiency parameter estimation is involved and the results of item parameter estimates are not affected, because the estimation of item and proficiency parameters are independent. Moreover, better equipped computer system can give faster computation for the parameter estimation.

The important aspect of the MCMC approach for parameter estimation of IRT models is the reasonable estimation accuracy and stability for the estimates. Simulation study can have a straightforward comparison between the estimates and the true parameters, which are available before the estimation. The accuracy of item parameter estimates increases as sample size increases, but decreases as the number of dimensions increases.

The estimation accuracy for item parameters can be seen from the comparison between the true and the estimates directly, which are presented in the RMSE tables (e.g., Table 3.8, 3.12, 3.13, 3.19, and 3.20) and plot figures (e.g., Figure 3.4 through 3.6 and Figure 3.10 through 3.13) in Chapter 3 for various simulation conditions. For

the unidimensional case, the item parameter estimates for both tests (e.g., the 30-item test and the 45-item test) are listed in Table 3.4, Table 3.6 and 3.7 for sample size 2000 and 5000 along with the standard errors. For multidimensional model parameter estimation, each item parameter estimate is not listed in a table but is plotted with the corresponding true parameters. The small difference between the true and the estimates of the item parameters indicates reasonable estimation. One can see in Table 3.4 and Table 3.6 and 3.7 on the item parameter estimates for unidimensional case, most of the absolute differences between the true and estimates are less than .1 and many of the standard errors of estimates are also less .1. More results are found in the summary statistics—RMSE. For unidimensional case, the RMSE for  $a$  parameters is less than .15 and arrives .07 when the sample size increases to 5000 (Table 3.8). The RMSE for  $b$  parameter is less than .11 and  $c$  parameter less than .05. For parameter estimation in multidimensional case, the RMSE is generally higher than the RMSE in the unidimensional models. For example in Table 3.12 and 3.13 for the RMSE for 3-dimensional model estimation, the RMSE for each  $a$  parameter estimates is generally higher than RMSE for  $a$  parameter in the unidimensional case; the RMSE in 5-dimensional item parameter estimation (Table 3.19 and 3.20) are in general higher than both the unidimensional and 3-dimensional case. One can conclude that as the dimension of proficiency increases in the model, the RMSE for item parameter estimates become larger, indicating poorer item parameter recovery. One simple interpretation to this observation is that the number of parameters to be estimated increases substantially as the proficiency dimension increases. Given the

same data structure and information, the more parameters need to estimate (as in the 3-dimensional and 5-dimensional model), the less information that the data contains for parameter estimation, and thus the less accurate the item parameter estimates. It is expected that the RMSE are larger in the 5-dimensional models than those in the 3-dimensional or unidimensional model. The good recovery of the item parameters can also be found from the plots of the true item parameters versus the estimates (e.g., Figure 3.4 through 3.6, Figure 3.10 through 3.13). In these figures the plots are closely around the reference line, indicating good estimates are obtained.

The precision of the proficiency estimates are assessed in terms of the correlation and plots of the true proficiency parameters versus the estimates. Large correlations are obtained for longer test (the 45-item test), but lower correlations are associated with higher dimensional tests (e.g., 5-dimensional test). The proficiency covariance matrix has negligible effects on proficiency parameter estimation.

The correlation tables show the correlations between the true proficiency parameters and estimates in terms of the number of dimensions, the sample size, and the test length (e.g., 3.9, 3.14, 3.15, 3.21 and 3.22 ). One can find that the correlations for the unidimensional case are the highest, more than .95 for every conditions in the simulation studies (Table 3.9). The correlations for the multidimensional cases (e.g., 3 dimensions and 5 dimensions) are generally lower than those in the unidimensional models, around .8 ~ .93 for each proficiency dimensions. The plots of the true proficiency versus the estimates in Figure 3.4 through 3.6 show the estimates are closely around the reference line for unidimensional model. However, the plots on Figure 3.10

through 3.13 for the multidimensional proficiency cases show the estimates relatively spread out from the line. The possible reason to explain the relations of the correlations with the proficiency dimensions is concerning the information that is contained in the data. One can expect better proficiency estimates or higher correlations for the lower dimensional models, in particular for the uni-dimensional model, because less parameters are required to be estimated in the same size of data structure and more information contained in the data is provided for the proficiency estimation. Better proficiency estimates is expected for longer tests if the higher dimensional model is used.

The estimation accuracy for both item and proficiency parameter estimates by the MCMC approach is clearly seen by the comparison of the results with the results from other procedures. For example, for unidimensional case, item parameter estimates in the 30-item test are calibrated from the standard procedure — MML/EM in BILOG-MG3, which is shown in Table 3.5. The results from the two approaches are comparable. However, the MCMC procedure, although from a Bayesian perspective, is flexible and convenient for much more complex IRT models. Furthermore, as Patz and Junker point out, one advantage of the MCMC procedure over traditional method is that this procedure is capable to estimate the exact joint posterior distribution for the parameters (Patz and Junker, 1999a).

The accuracy of the estimation by MCMC is clearly seen from the consensus estimation on the replication of data sets and the consensus estimation on the replication of multiple chains. This is also the aspects of the stability of the parameter estimation

of the MCMC approach. It is seen from Table 3.3 for the unidimensional case, the three independent chains yield very stable estimates of item parameters for the 30-item tests. Similar results are obtained for the 45-item tests and higher dimensional model parameter estimation. For the same data set, parameter estimates are stable from three independent chains with different initial values indicating the posteriors of the model parameters reach the stationary status. That is why the parameter estimates do not depend on the initial values. The item parameter estimates are not only stable across the multiple chains, but also stable across data sets (e.g., Table 3.8, 3.9).

It seems difficult to increase the estimation precision for both item and proficiency parameter estimates in IRT models at the same times. When the sample size increases for a fixed number of items in a test, the item parameter estimates are expected to be improved. For a fixed group of examinees, the proficiency parameter estimates are expected to improve as the number of items in a test increase. One can argue that for a fixed number of items in a test, the number of item parameters to be estimated is fixed and increasing the sample size of examinees provides more information for estimating item parameters. Therefore, the standard error of estimates decreases. When estimating proficiency parameters for a fixed number of examinees, the number of proficiency parameters to be estimated will be improved as the number of items increases in the test, because the test provides more information for estimating proficiency parameters. This also happens to the parameter estimation using the MCMC procedures. It is seen from Table 3.13 and 3.14 that for a fixed test (e.g., the 30-item

test or the 45-item test), item parameter estimates get better in terms of RMSE when the sample size changes from 2000 to 5000.

The proficiency covariance is well recovered in the MCMC procedure and the estimation of the proficiency covariance matrix does not affect the item parameter estimates.

The relations between the estimates and the design variables for a test (e.g., the test length, the sample size of examinees, and the number of dimensions) are helpful for suggesting a general guideline for parameter estimation. For example, to require accurate item parameter estimates for the unidimensional model assuming perfect model-data fit, if a test consists of 30 items, the number of 2000 examinees is good enough. But with the same number of 30 items for estimating item parameters from the 3-dimensional model, more than 2000 examinees (e.g., 5000) could achieve the estimation precision. Similarly, for the 5-dimensional model, more than 5000 examinees (e.g., 8000 or more) could help to reach the same estimation precision. For proficiency estimates using the unidimensional model, the number of 30 items can provide reasonable good estimation, as seen in the correlation Table 3.9 and plot Figure 3.3. But for the 3-dimensional test, the number of 45 items could provide reasonable good estimation for proficiency estimates, as seen in Table 3.14 and 3.15 and Figure 3.17. For the 5-dimensional model, more than 45 items (e.g., 60 items) could help for reasonable good proficiency estimation.

One limitation for the MCMC approach estimating multidimensional IRT model parameters except the extensive computation, is the number of dimension is given.

But the number of dimensions is not generally available in real data analysis. How would the performance of the MCMC approach be if the number of dimension is less or more than that of the required dimensions in the test? This is an interesting practical issue and worthwhile for further research efforts. This issue is in fact also a model-data fit issue rather than parameter estimation issue (the focus of the whole research), or sensitivity issues on parameter estimation using the MCMC approach. The reality is the estimates are acceptable on the basis of the model-data fit. However, the MCMC approach does not give any mechanism to diagnose whether or not the data fit the estimating model. How much additional errors would be introduced because of the model-data having not adequately fit? This practical issue would give challenges to the MCMC estimation.

In the simulation studies, the proficiency covariance matrix varied from a special pattern (e.g., all off-diagonal elements are the same) to a general one and the effects of the proficiency covariance matrix on the parameter estimation are carefully examined, the proficiency population is assumed from multivariate normal or standard normal. If the examinee groups are not from a normal distribution, does the approach still yield accurate and stable estimation? This issue also deserves further research efforts, because the examinees might not come exactly from a normal population in many applications.

In addition, the metric for both item and proficiency parameters is established by a well-defined set of anchor items, which are often placed in the first positions in the tests. The anchor items help with solving the indeterminacy problems that

is inherited in many IRT models. However, the choice of anchor items are often subjective, and therefore may influence the establishment of the proficiency scales. Further research is needed to investigate the effects of the anchor items on parameter estimation using the MCMC approach. In real data applications, how can one choose a useful set of anchor items that help with the model identification and meanwhile ensure accurate parameter estimation?

Finally, the item parameter estimates by MCMC methods are compared with the estimates by TESTFACT, and the results show that the estimates from MCMC methods are better than those from the TESTFACT. Table 4.1 shows the item parameter estimates by TESTFACT for the 30-item test with 3 dimensions to 2000 examinees (i.e., the first replication of the data). Table 4.2 shows the item parameter estimates by TESTFACT for the 30-item test with 5 dimensions to 2000 examinees (i.e., also the first replication of the data). The input of the estimates for the pseudo-guessing parameters is the true values for the  $c$  parameters. Compared with the true item parameters (Table 3.10 and Table 3.16) and the estimates by MCMC (Table Table 3.13, 3.19, and 3.20), the item parameter estimates by TESTFACT in Table 4.1 and 4.2 in general seem a little bit worse. In addition, the results from TESTFACT have some deviant values (e.g., item 11, item 17, item 18, item 19, item 21 in Table 4.2 for 5-dimensional case).



Table 4.1: TESTFACT Item Parameters estimates for 30-Item Test (Dim = 3)

Item	$a_1$	$a_2$	$a_3$	$d$
1	1.25	-0.2	-0.17	-0.27
2	-0.09	0.65	-0.16	0.02
3	-0.45	-0.37	2.28	-1.2
4	1.97	-0.33	-0.35	0.53
5	0.86	-0.21	-0.13	0.22
6	1.37	-0.22	-0.21	1.47
7	0.65	-0.22	-0.08	1.57
8	-0.55	2.64	-0.41	-0.53
9	-0.27	1.86	-0.41	-1
10	-0.11	0.67	-0.15	-0.6
11	-0.16	1.08	-0.15	1.16
12	-0.29	-0.34	1.47	-0.33
13	-0.24	0.04	0.89	2.22
14	-0.03	-0.32	1.31	-1.46
15	-0.28	-0.36	1.83	-0.56
16	0.44	1.54	1.26	-1.48
17	1.11	1.78	0.19	-0.24
18	0.3	0.71	1.19	0.86
19	1.28	1	1.17	0.21
20	0.04	0.48	1.5	0.95
21	2.06	0.4	0.53	-1.39
22	1.29	0.75	1.96	-2.2
23	0.8	1.16	0.78	-0.42
24	2.27	0.19	-0.09	-1.55
25	1.35	1.45	0.45	-1.91
26	0.58	0.98	0.44	0.21
27	0.61	0.91	1.02	0.66
28	1.13	0.77	1.31	0.37
29	0.95	0.41	1.62	0.16
30	0.47	1.99	1.09	-0.87

Table 4.2: TESTFACT Item Parameters Estimates for 30-Item Test (Dim = 5)

Item	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$d$
1	0.88	-0.07	-0.29	-0.11	-0.02	1.59
2	-0.54	3.67	-0.59	-0.43	-0.66	-0.76
3	-0.05	-1.67	7.62	-0.88	-1.03	-1.84
4	-0.16	-0.4	-0.4	1.71	-0.25	1.48
5	-0.21	0	-0.25	-0.38	1.47	-0.58
6	1.96	-0.27	-0.39	-0.35	-0.24	-0.06
7	2.67	-0.31	-0.33	-0.54	-0.51	-0.75
8	-0.06	0.83	-0.14	0.03	-0.21	-0.5
9	-0.17	2.43	-0.38	-0.42	-0.42	1.17
10	0.14	-0.2	1.3	-0.16	-0.25	-1.24
11	-22.14	-5.89	98.05	-15.65	-22.07	-76.86
12	-0.27	-0.22	-0.31	1.89	-0.41	1.66
13	-0.15	-0.16	-0.25	1.44	-0.31	2.13
14	-1.17	-2.15	-1.93	-1.48	10.28	2.83
15	0.01	-0.11	-0.14	-0.08	0.45	-0.42
16	3.19	-0.12	1.26	1.25	-0.51	0.3
17	15.39	1.78	26.26	7.73	26.4	-35.36
18	-0.62	6.8	0.43	6.35	10.72	7.27
19	-4.07	3.32	5.28	8.76	9.89	14.02
20	0.36	-0.06	1.26	1.74	1.24	1.57
21	-5	2.5	15.72	43.88	8.1	24.18
22	1.23	0.66	2.01	0.1	1.68	1.94
23	0.78	0.2	0.07	0.63	-0.1	0.72
24	0.81	2.59	1.17	-0.29	-0.03	0.21
25	2.02	-0.18	0.36	0.15	0.82	0.1
26	1.59	0.07	1.14	0.42	0.56	-0.26
27	0.7	0.21	0.05	0.45	0.03	0.12
28	1.13	0.25	0.55	-0.09	-0.02	-0.48
29	0.87	0.35	-0.05	0.48	0.21	0.25
30	0.3	0.28	0.74	0.52	0.5	0.5

# Bibliography

- [1] Ackerman, T. A. (1990). An evaluation of the multidimensional parallelism of the EAAP Mathematics Test. *Paper presented at the Meeting of the American Educational Research Association*, Boston, MA.
- [2] Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement* 29(1), 67-91.
- [3] Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, 17, 251-269.
- [4] Baker, F. B. (1990). Some observations on the metric of BILOG results. *Applied Psychological measurement*, 14, 139-150.
- [5] Beguin, A. A., Glas, C. A. W. (1998). ED428100. MCMC Estimation of Multidimensional IRT models. Research Report 98-14.
- [6] Besag, J., Green, P. J., Higdon, D. M., and Mengersen, K. L. (1995). Bayesian Computation and Stochastic Systems (with discussion). *Statistical Science* 10, 3-66.
- [7] Birnbaum, A. (1957). Efficient design and use of tests of a mental ability for various decision-making problems. (Series Report No. 58-16. Project No. 7755-23). USAF School of Aviation Medicine, Randolph Air Force Base, Texas.
- [8] Birnbaum, A. (1958a). Further considerations of efficiency in tests of a mental ability. Technical Report No. 17. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas.
- [9] Birnbaum, A. (1958b). On the estimation of mental ability. Series Report No. 15. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas.
- [10] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397-472). Reading, MA: Addison-Wesley.

- [11] Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*.
- [12] Bock, R. D., Gibbons, R., & Muraki, E. J. (1988). Full information item factor analysis. *Applied Psychological Measurement*, 12, 261-280.
- [13] Carlson, J. E. (1987). Multidimensional item response theory estimation: A computer program (Research Report ONR 87-2). Iowa City, IA: The American College Testing Program.
- [14] Bock, R. D. and Lieberman, M. (1970). Fitting a response model for  $n$  dichotomously score items. *Psychometrika*, 35, 179-197.
- [15] Bolt, D. M. and Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychomological Measurement*, 27(6), 395-414.
- [16] De-la-Torre, J, Patz, R. J. (2001). ED 464143. Item Response Theory Equating Using Bayesian Informative Priors. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education* (Seattle, WA, April 11-13, 2001).
- [17] Embreston, S. E. and Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- [18] Fox, J. P. (2002) Multilevel IRT Using Dichotomous and polytomous Response Data. Research Report.
- [19] Fraser, C. (1988). NOHARM II. A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, Australia: The University of New England, Center for Behavioral Studies.
- [20] Gamerman, D. (1997). *Markov Chain Monte Carlo*. New York: Chapman & Hall.
- [21] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Second Edition. Chapman & Hall.
- [22] Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Sciences*, 7(4), 457-472.
- [23] Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restroation of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 6, 721-741.
- [24] Gilks, W. R., Richardson, S., and Spiegelhalter , D.J., eds. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

- [25] Gill, J. (2002). *Bayesian methods for the social and behavioral sciences*. Chapman & Hall/CRC.
- [26] Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- [27] Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: Principle and Applications*. Kluwer Nijhoff Publishing.
- [28] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.
- [29] Hulin, C. L., Lissak, R. L., and Drasgow, F. (1982). Recovery of two and three parameter logistic item characteristics curves: A monte Carlo study. *Applied Psychological Measurement*, 6, 249-260.
- [30] Kiefer, J., and Wolfowitz, J. (1956). Consistency of maximum likelihood estimates in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27, 887-890.
- [31] Kim, S. H., & Cohen, A. S. (1998). ED420689. An Evaluation of a Markov Chain Monte Carlo Method for the Two-parameter Logistic Model.
- [32] Lemann, E. L., Casella, G. (1998). *Theory of point estimation*. Second edition. Springer-Verlag New York, Inc.
- [33] Li, J. C., Woodruff, D. J. (2001). ED 462419. Bayesian Statistical Inference for Coefficient Alpha. ACT Research Report Series.
- [34] Little, R. J. A., and Rubin, D. B. (1983). On jointly estimating parameters ad missing data by maximizing the complete-data likelihood. *The American Statistician*, 37, 218-220.
- [35] Lord, F. (1952). A theory of test scores. *Psychometric Monograph*, No. 7.
- [36] Lord, F. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13, 517-548.
- [37] Lord, F., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- [38] Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- [39] Maris, G. & Maris, E. (2002). A MCMC-Method for Models with Continuous Latent Responses. *Psychometrika Vol. 67, No. 3*, 335-350.
- [40] Matthews-Lopez, J. L., Hombo, C. M. (2001). ED 454268. Modeling the Hyper-distribution of Item Parameter to Improve the Accuracy of Recovery in Estimation Procedures.

- [41] McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric monographs*, No. 15.
- [42] McDonald, R. P. (1985). Unidimensional and multidimensional models for item response theory. In D. J. Weiss (Ed.), *Proceeding of the 1982 Computerized Adaptive Testing Conference* (pp. 127-148). Minneapolis: University of Minnesota, Department of Psychology, Psychometrics Methods Program.
- [43] McKinley, R. L., & Reckase, M. D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods & Instrumentation*, 15, 389-390.
- [44] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21, 1087-1092.
- [45] Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115-132.
- [46] Neyman, J., and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, 16(1), 1-32.
- [47] Patz, R. J. Junker, B. W. (1999a). A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models. *Journal of Educational and Behavioral Statistics*, 24(2), 146-178.
- [48] Patz, R. J. Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24(4), 342-366.
- [49] Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- [50] Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.
- [51] Reckase, M. (1996). A linear logistic multidimensional model for dichotomous item response data. In W. Van der Linden, & R. Hambleton (Eds), *Handbook of modern item response theory* (pp.271-286). New York: Springer - Verlag.
- [52] Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Unidimensional data from multidimensional testes and multidimensional data from unidimensional test.
- [53] Reckase, M. D. & Hirsh, T. M. (1991). Interpretation of number-correct scores when the true number of dimensions assessed by a test is greater than two.

*Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.*

- [54] Roussos, L. A. (1995). A new dimensionality estimation tool for multiple-item tests and a new DIF analysis paradigm based on multidimensionality and construct validity. *Unpublished doctoral dissertation*, University of Illinois at Urbana-Champaign.
- [55] Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No. 17.
- [56] Samejima, F. (1972). A general model for free-response data. *Psychometric Monograph*, No. 18.
- [57] Segall, D. O. (2001). General ability measurement: an application of multidimensional item response theory. *Psychometrika*. Vol. 66, No. 1, 79-97.
- [58] Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*. Vol. 61, No. 2, 331-354.
- [59] Stegelmann, W. (1983). Expanding the Rasch model to a general model having more than one dimension. *Psychometrika*, 48, 259-267.
- [60] Tierney, L. (1991). Exploring Posterior Distributions Using Markov Chains. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. E. M. Keramidas (ed.). Fairfax Station, VA: Interface Foundation. pp. 563-570.
- [61] van der Linden, W. J., & Hambleton, R. K. (1996). *Handbook of modern item response theory*. New York: Springer.
- [62] Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45, 479-494.
- [63] Williamson, D. M., Johnson, M. S., Sinharay, S., & Bejar, I. I. (2002). ED 464948 Hierarchical IRT Examination of Isomorphic Equivalence of Complex Constructed Response Tasks.
- [64] Wright, B. D., & Stone, M. H. *Best test design*. Chicago: MESA, 1979.
- [65] Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. Recovery of Item Parameter in the Nominal Response Model: A Comparison of Marginal Maximum Likelihood Estimation and Markov Chain Monte Carlo Estimation. *Applied Psychological Measurement*, 26(3), 339-352.

MICHIGAN STATE UNIVERSITY LIBRARY



3 1293 02736 8