Bi-factor Multidimensional Item Response Theory Modeling for Subscores Estimation, Reliability, and Classification

By

Zairul Nor Deana Md Desa

Submitted to the Department of Psychology and Research in Education and the
Faculty of the Graduate School of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Committee members

William P. Skorupski, Ph.D.,
Chairperson

Neal Kingston, Ph.D.

Bruce B. Frey, Ph.D.

Vicki Peyton, Ph.D.

Carol M. Woods, Ph.D.

Date defended: ______________________

The Dissertation Committee for Zairul Nor Deana Md Desa certifies
that this is the approved version of the following dissertation :

Bi-factor Multidimensional Item Response Theory Modeling for Subscores Estimation,
Reliability, and Classification

William P. Skorupski, Ph.D., Chairperson

Date approved:

Abstract

In recent years, there has been increasing interest in estimating and improving subscore reliability. In this study, the multidimensional item response theory (MIRT) and the bi-factor model were combined to estimate subscores, to obtain subscores reliability, and subscores classification. Both the compensatory and partially compensatory MIRT models are defined with bi-factor structure. A Monte Carlo study with 1,500 examinees was carried out for each model to examine two different test lengths (30 and 60 items) and five levels of item discrimination between primary and specific abilities (.50, .75, 1.0, 1.25, 1.50). The Markov Chain Monte Carlo (MCMC) with the Gibbs sampling method was applied to simultaneously estimate the expected a posteriori (EAP) subscores for primary and specific ability dimensions. Results were evaluated in light of estimation accuracy and fit, subscore reliability based on the Bayesian marginal reliability, and subscore classification based on subscore separation index. Despite a very minimum computing intensity for the MCMC simulation, both bi-factor compensatory and bi-factor partially compensatory models produced higher subscores reliability resulted from lower bias and reduction in the error variance of EAP subscores in all ability dimensions. These improved subscores reliability that also arrived at a higher discrimination level and for a longer test. This study found the bi-factor compensatory model to show better potential in classifying the magnitude of distinction between specific abilities and primary ability. Whereas, the bi-factor partially compensatory minimized the classification of subscores between the specific and primary abilities.

Acknowledgements

I would never have been able to complete my dissertation without the credible guidance from all of the committee members, supports from friends, and family members. I feel indebted to many people who have greatly inspired and supported me during my several years of my Ph.D. study at the University of Kansas. I would like to express my deep gratitude to my advisor, Dr. William P. Skorupski, for his excellent guidance, patience and providing me with motivational thoughts for doing this research. I would like to thank Dr. Neal Kingston, Dr. Bruce B. Frey, Dr. Vicki Peyton, and Dr. Carol M. Woods for their heartful encouragement, academic stimulus, and generous help throughout my study in the program. I also want to thank Dr. Kristopher J. Preacher who served on my Comprehensive Exam Committee, and Dr. Paul E. Johnson for devoted so much time and effort into my study and his patience with my questions. I will not forget enjoyable discussions and thoughts in academia with these people. My gratitude also goes to friends and colleagues for their ideas, comments, support and many cherished moments. Also, I am indeed obliged to the Universiti Teknologi Malaysia and the Ministry of Higher Education Malaysia for giving me the opportunity and financial support to study and carry out research abroad from 2008 to 2012. I owe permanent love from my amazing and precious parents, dear husband and daughter for their understanding and fascinating support. My infinite loop of deepest gratitude and love especially goes to them.

Contents

List of Figures

List of Tables

List of Equations

Chapter 1

Introduction

Test scores can provide an informative inference about examinees performance on an assessment. Reliable test scores are reported for the acquisition of a broad range of content materials that is fair enough for the examinees to be compared with the respective norm groups. This is important for test validation, which requires broad content coverage. Standard 5.12 in the 1999 Standards for Educational and Psychological Testing AERA et al. (1999) emphasizes that scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established. However, reporting subscores of test subscales is permissible for reliable scores over the total scores, indicating significant consequences for the examinees and the stakeholders.

1.1 Background of the Study

Judicious reviews of psychometric properties provide for an important mechanism to the test theory as well as reduce any unintended social consequences (Messick, 1995) associated with the use of subscores, especially the subscores that are tied to explicit or implicit objectives of instruction or to specific content domains. The implications of reporting subscores are also addressed under the No Child Left Behind Act of 2001 NCLB (2001) that every state is required to develop an accountability assessment system to measure statewide progress and evaluate school performance. The NCLB Act of 2001 affirming that:

> Such assessments shall produce individual student interpretive, descriptive, and diagnostic reports . . . that allow parents, teachers, and principals to understand and address the specific academic needs of students, and include information regarding achievement on academic assessments aligned with State academic achievement standards, and that are provided to parents, teachers, and principals as soon as is practicably possible after the assessment is given, in an understandable and uniform format, and to the extent practicable, in a language that parents can understand. (NCLB, Part A, Subpart 1, Sec. 2221[b]3[C][xii], 2001)

Concentration on the precise scoring methods for the specific content domains in which the examinees may be having difficulties is needed for informed feedback. Therefore, sophisticated scoring procedures for estimating the subscores can lead to more reliable and valid scoring reports about examinee's strengths and weaknesses in content-related subjects. Thus, reliable and valid scores are very important for accountability assessments that might be aided to optimal design for diagnostic purposes.

It is very common for test scores from large-scale assessments to be calibrated from unidimensional space and reported as item response theory (IRT) scale scores. Traditional IRT scale scores or IRT subscales scores might have made use of the summed scores or as a measure the true proficiency (Hambleton & Jones, 1993; Hambleton & Swaminathan, 1985; Lord, 1980, 1952). Summed scores or number-correct scores were judged unacceptable and unattractive features for some of the stakeholders where the scores were reported to intense public scrutiny in large-scale testing.

The property of summed scores based on classical test theory does not allow that examinees with the same number of correct items received the same scale scores on a test. Thus, subscores from item-patterns are used instead to stabilize the subscale scores (Wainer et al., 2001). Subscores may be used to infer an examinee's strengths and weaknesses on the subscales of specific abilities (e.g. algebra, geometry, and number sense in mathematics), and are important for projective performance. Subscores may be computed from the examinee's response pattern from each subscale item, and the probability of success from each item within the subscale is predicted to give the total score, in turn, that infers the overall examinee's ability based on the rest of the test (Yen, 1987).

1.1.1 IRT Scale Score

Both unidimensional IRT (UIRT) models and multidimensional IRT (MIRT) models developed from probabilistic IRT models and are straightforward to understand. UIRT and MIRT models are capable of predicting examinee proficiency or cognitive abilities, and resolve the classical test theory (CTT) limitations or incapabilities, given the examinees responses on test items. Items in Large-scale assessments (e.g. SAT, ACT, GRE, TOEFL, TIMSS, PISA, NAEP and PIRLS) are designed to measure proficiency in multiple or more specific content domains within a larger subject such as algebra and geometry within mathematics, and physical and life sciences within science. Test items from large-scale assessments are typically calibrated based on unidimensional IRT models due to their time consuming nature and demands from testing companies and stakeholders. Moreover, with unidimensonal IRT, there is very limited information on examinee performance on the specific content domains because the reported score is based on a single dimension which covers only the broad macroskills. Thus, student achievement on the specific content domains are barely interpreted or evaluated.

IRT scale scores from UIRT or MIRT models provide rich information about item characteristics and distribution of examinees' performance on each item continuum of their ability metric scale that typically on Normal distribution with zero mean and standard deviation 1.0 (Hambleton & Swaminathan, 1985; Thissen, 1982). IRT scale scores have many applications as items and examinees can be placed on a common scale, for example, test linking, equating and scaling (Kolen & Brennan, 2004), item banking, computer-based testing and computerized adaptive testing (Drasgow & Olson-Buchanan, 1999; Mills et al., 2002; Parshall et al., 2001; van der Linden & Glas, 2000; Wainer & Dorans, 2000). Thus, the examinee's reported proficiency is more conducive for further psychometric analysis and enhances test score interpretations for multiple groups of examinees or multiple test forms. Also, for a test that is defined to have multidimensional structure, IRT scaled scores for all abilities as well as specific test objectives or skills can be estimated simultaneously and, therefore, be reported for many educational purposes such as student profiling and diagnostic feedback to teachers on student's strengths and weaknesses.

1.1.2 Bi-factor Score

Due to the strong relationship between the development of factor analysis and IRT, there has been recent attention to these sophisticated approaches to estimate scale scores for tests that have been defined as having more than one content domain or construct. The confirmatory-based IRT on the two-dimensional test called the bi-factor model has been in development since the 1990's based on a unidimensional IRT framework (Gibbons et al., 1990; Gibbons & Hedeker, 1992; Gibbons et al., 2007).

The bi-factor model was originally introduced as an extension of the Spearman's two factors, when a test measures a general factor as well as secondary factors that the secondary are sometimes treated as group factors (Holzinger & Swineford, 1937). This model is also known as the Thurstone simple-structure second order of the general factor, and more than one factor is included in the secondary factor (Swineford, 1941; Thurstone, 1947; Schmid Jr, 1957; Schmid & Leiman, 1957).

Therefore, it is certainly permissible for IRT scaled subscores to be estimated from item-patterns of two measures of ability or proficiency in both general (or primary) and secondary specific content domain. Pairing the bi-factor model and the IRT model allows a primary (or general) factor to be estimated from item responses, and that test items measure two or more mutually exclusive secondary factors. It is relevant that test items with specific content domain share a common targeted domain or process skill. As the variance accounted for by the specific content domains increases to explain examinee's unique abilities, the association between items within each specific content domain is expected to be greater than test items assessing the examinee's primary ability or process skill from the primary domain. Consequently, the bi-factor model may yield more accurate test score estimates than other psychometric models when a primary and specific content domain are considered.

1.2 Scoring Subscores Based on Item Response Theory

Subscores can be estimated in several ways. The basis behind the IRT estimation is to convey information provided by the characteristics of the IRT models when an examinee correctly responds to test items. In cases when the score pattern is zeros (incorrect) and ones (correct), the probability of combining the score pattern on the examinee's entire pattern of item responses portrays the joint probability between lows and highs that an examinee would respond negatively or positively. Highest probability of correct is obtained to describe the IRT scale score. This method is known as likelihood estimation method and is called the Maximum Likelihood Estimate (MLE) or joint Maximum Likelihood Estimate (JMLE) of examinee proficiency (Baker, 1985; Yen, 1984; Thissen, 1982).

However, these methods are tedious and involve a lot of computing when there are a large number of examinees and/or a large number of items in a test. Also, JMLE cannot be used when responses to test items in the primary and/or in the specific content domains are all correct (i.e. a perfect score) or responses to test items in the primary and/or in the specific content domains are all incorrect. The product of the likelihood will yield positive infiniteability for an all correct pattern, and tends towards negative infiniteability for an all incorrect pattern (Baker, 1985; Baker & Kim, 2004; Hambleton & Swaminathan, 1985; Thissen & Orlando, 2001). The same problems are observed with multidimensional estimation (Reckase, 1997, 2009). JMLE and MLE estimation methods have disadvantages when the calculation involves multidimensional integration over the ability parameter, and the number of calculations involved increases exponentially with the number of dimensions.

A better estimation approach for both the UIRT and the MIRT is the Marginal Maximum Likelihood Estimation (MMLE) sometimes referred to as the full-information maximum likelihood estimation method (Bock & Aitkin, 1981; Bock et al., 1988). This method is used to overcome inconsistent and insufficient estimation problems in MLE and JMLE. MMLE does not estimate ability parameter from the likelihood but is applied within a Bayesian framework after item parameters have been estimated. MMLE is available in the IRT software, for example BILOG-MG

(Binary Logistic Model for Multiple Group) and MMLE is explained in details in TESTFACT (Croudace et al., 2005).

Another estimation method based on the Bayesian approach is the Markov Chain Monte Carlo (MCMC). The MCMC is known as a fully Bayesian approach. MCMC includes Monte Carlo integration in the likelihood from the conditional probability of the IRT model using Markov chain procedures (Gilks et al., 1996). Monte Carlo integration draws samples from a target distribution for an item or examinee's parameter, then taking the averages to approximate their expectations. Markov Chain ensures that the chosen draws are independent between chains, and the target distribution is at its stationary distribution. Thus, the estimated values of the item and ability parameters are both stable and invariant measures.

To reach a stationary distribution, a sequence of random events happens at $t+1$depends only on the state at time t and not at any prior state. That is, the probability of moving to a new state is conditional on the current state. Stationary solution (i.e. stable or optimal parameter estimates) will behave approximately like samples from the target distribution of interest. The general MCMC estimator is called the Metropolis–Hastings Algorithm (MH algorithm) (Metropolis, 1987; Hastings, 1970) and the most straightforward method is the Gibbs sampling (Geman & Geman, 1984). These two MCMC methods give marginal posterior distributions of the parameters of interest.

MMLE and MCMC methods are fairly slow but this is not an issue. Computer speeds are increasing very rapidly. One problem with the used of MMLE in estimating item parameters from the MIRT models is that an issue needs to deal with indeterminacies for unique solution in the multiple dimensions of ability parameter. A solution to the indeterminacies problem of item parameter estimations can be achieved by fixing the the ability parameter to be estimated from multivariate normal distribution with mean vector zero elements and an identity matrix for the variance-covariance matrix (Reckase, 2009). Whereas, common issues in MCMC are the specification of the prior distribution and convergence diagnostic. Examples for prior selection are choosing conjugate prior and highly proper noninformative prior that would ensure the posterior distribution is returned from the same distribution, and these types of priors are more computationally efficient.

Results from the MCMC algorithm should be carefully evaluated so that the estimated item and ability values are truly representative the underlying stationary solutions.

1.2.1 Improving Reliability of Subscore Estimates

Many studies have been proposed since the late 1980's to ensure high subscore reliability for clustered test items on specific content domains. Yen (1987) proposed the objective performance index (OPI) that is accomplished through a Bayesian IRT procedure. The OPI is utilized to produce a more stable score which is use to report scores that are separated by their test objectives. The OPI score that is based on the posterior distribution is observed to be sufficiently accurate so as to be useful for score reporting. Teachers have used the score to infer their students' strengths and weaknesses in a particular test objective.

Subscores on specific content knowledge can also be predicted based on regression technique on subscores or Kelly's regression for true score estimate (Kelly, 1927, 1947). Wainer et al. (2001) proposed a method called subscores augmentation which is achieved by regressing any particular subscores (e.g. algebra) on any other subscores (e.g. geometric and trigonometry). The study also implemented a multivariate empirical Bayes estimate (EB) based on the correlation among the subcales that posit associations between the subscores being observed. The performance of the EB subscores augmentation method has been further studied in a simulation study by Edwards (2002) Edwards & Vevea (2006) and studied empirically in Skorupski & Carvajal (2009) on statewide testing data. All those studies came to the same conclusions, for the test that is unidimensional in nature the proposed subscores stabilization is substantially improves the precision of the scores and produces a reduction in error of estimates as well as outperform traditional subscores (i.e. number-correct scores).

Other studies on improving reliability of the subscore estimates have been carried out by incorporating ancillary or collateral information of other subtests information in a test into the estimation of the ability as well as item parameters. Ancillary or collateral information is a method of taking the information of the correlations between multiple abilities being measured in a test and to infer about examinee proficiency only in the overall performance or only in the specific domain abilities. Similar to OPI and empirical Bayes subscores stabilization techniques, collateral information is accomplished in a sense of using items in other test objectives into a calibration of the particular item in one objective (Ackerman & Davey, 1991; Bock et al., 1997; Pommerich et al., 1999; Shin, 2007; Tao, 2009). Thus, all other test subscales are contributing to measuring one subscale, and the estimated subcsores showed greater accuracy.

Subscores are equally important for reliable scores that are associated with multiple dimensions of test batteries. As specific content area are highly correlated with each other, then reporting subscores is important for instructional and remedial decisions at student and school-level decisions (Sinharay et al., 2007; Haberman, 2008; Haberman et al., 2009; Puhan et al., 2010).

The application of Bayesian approach has shown improved reliability of the overall score or subscores estimates (Kolen & Tong, 2010; Wainer et al., 2001; Yen, 1987; Skorupski & Carvajal, 2009; Tao, 2009; Haberman, 2008; Sinharay et al., 2007; Ackerman & Davey, 1991; Mislevy, 1987). For example, Skorupski & Carvajal (2009) and Wainer et al. (2001) observed substantial improvement in reliability of the subscore estimates when the Bayesian approach was used to stabilized scores from one subtest by borrowing information from other subscores. This method is called the subscrores augmentation method.

1.3 Statement of the Problem

Successful studies in improving subscore reliability based on the CTT and the UIRT using information from other subtests or subdomains have raised interest in the field of psychometrics and educational measurement so as to improve subscore estimation using the MIRT models. Empirical and simulation studies have observed that the MIRT models demonstrate better improvement of subscore reliability over the CTT and UIRT approaches, and the estimated subscores provide added value in addition to the total score (DeMars, 2005; Kahraman & Thompson, 2011; Haberman & Sinharay, 2010; Sheng & Wikle, 2009; Yao, 2010; Thissen & Edwards, 2005).

Since its development that started almost two decades ago, improving subscore estimation for the multidimensional IRT is still in its infancy stage compared to unidimensional IRT approaches. Also, only the compensatory MIRT model, in which one ability can be offset by an increase in other abilities, has been considered to explain examinees' process skills on tests with multidimensional structures. Tate (2004) studied the implication of multidimensional test structure on subscore estimates. The simulation study was to compare MLE and expected a posteriori (EAP) approaches to estimate subscores. Many of the aforementioned researchers have applied MCMC to estimate subscores. In summary, despite the issue of inefficient computing time, the studies found that Bayesian estimation techniques such as the MCMC method tends to produce subscore estimates that is feasible for test scoring and reporting. Most of the studies found that the estimated subscores based on the MIRT models are slightly better than other estimation methods from the classical test theory and the unidimensional IRT techniques.

Scoring subscores from tests with a truly multidimensional structure are emergent, but are in early development , and improving reliability of subscores from perspective of the unidimensional IRT is promisingly extended. The idea of using the empirical Bayes estimate from the unidimensional IRT in Wainer et al. (2001) is observed to have improved reliability on the estimation of subscores. This is performed by augmenting each subscore from all other subscores in a test with informed priors and the precision of an estimated subscore can be improved by shrinking the estimate towards the average value so that the empirical Bayes estimates would be very close to the reliable true observed score.

For the bi-factor model in context of the MIRT application, the empirical Bayes estimate based on the Bayesian expected a posteriori (EAP) can be used to estimate the IRT scale subscores. As the purposes of the test entails measuring primary ability as well as specific content domain abilities, subscores from EAP technique can also be estimated. Unlike MLE or JLMLE methods, for the cases of all-correct and/or all-incorrect answers from a test, the EAP subscores for an examinee will not go to infinite values. Essentially, this method incorporates likelihood function of the item score and prior distributions from examinee parameters to estimate posterior distributions of all elements of subscores. Empirical Bayes estimates can be applied for each subscore when it is estimated from the EAP estimation method. The EAP estimation method shrinks the subscores linearly toward the mean for the population and the estimated subscores are proportional to their standard errors, $SD[\theta]$. Thus, inferences about examinees proficiencies for both primary and specific content domains can be made based on the posterior distributions.

In addition, existing literature shows effective application of the bi-factor model to clinical psychology and health related outcomes (see Gibbons et al., 2007; Reise et al., 2011, 2007). Very limited number of studies, however, have used the bi-factor model with the application from the MIRT model estimation specifically from the compensatory and partially compensatory models and almost none provide supporting evidence for subscore estimation via the partially compensatory model. In the compensatory MIRT model, deficiency in one ability can be offset by an increase in other abilities. On the other hand, in the partially compensatory MIRT model, deficiency in one ability cannot be offset by an increase in other abilities but sufficient levels of each measured ability are required (Ackerman et al., 2003; Bolt & Lall, 2003; Reckase, 1997, 2009).

The bi-factor models are not well established in the field of educational measurement. The studied bi-factor models in educational outcome are more prevalent as a testlet-based MIRT model (Bradlow et al., 1999; DeMars, 2006; Li et al., 2006; Rijmen, 2009, 2010) or as a hierarchical second-order IRT model (Immekus & Imbrie, 2008; Reise et al., 2007; Rijmen et al., 2008; Yung et al., 1999).

Few studies exist that provide limited information on the subscore estimation from educational assessment based on the MIRT framework (DeMars, 2006; Edwards & Vevea, 2006). Both studies are simulation studies, and closely related to testlet-based approach as the true model for tests were designed to have items within reading passages, thus measurement of the primary and specific proficiency on content domains of other subject areas was not emphasized. Also, only the compensatory multidimensional IRT has been considered in the parameter estimation, but there are also mental tests that require partially compensatory activities to respond to test items (e.g. reading comprehension that requires both reading comprehension and decoding skills). Unpublished empirical paper by DeMars (2005) observed the bi-factor model within the MIRT framework but the proposed model was not clearly specified. The implication of the application of the bi-factor model with the existing multidimensional IRT models in subscore estimation is not fully developed or studied. Therefore, further studies in this direction can be promising for improving reliability of subscores estimation. Moreover, solutions about the accuracy of the parameter estimations from the bi-factor MIRT models can be examined.

Not many studies have focused on the Bayesian approach to improving subscore estimations for a test with multidimensional structure as well as to infer about examinees primary and specific abilities. Recent studies (e.g. de la Torre et al. 2011; Edwards & Vevea 2006; Rijmen 2010; Tao 2009; Yao 2010; Babcock 2011) have revealed very limited amount of information on the bi-factor model with the confirmatory compensatory and the confirmatory partially compensatory multidimensional IRT. Consequently, none of the studies have provided sound methods for simultaneous estimation of the primary and specific domain abilities using the bi-factor confirmatory compensatory and the bi-factor confirmatory partially compensatory MIRT models to improve the precisions of the subscore estimates.

1.4 Purpose of the Study

Given the need for improving reliable subscores, and a paucity of attention to estimating subscores based on the Bayesian estimation from the confirmatory compensatory and the confirmatory partially compensatory multidimensional IRT models that are paired with the bi-factor model, this study proposes two models to improve the reliability of subscore estimates from the bi-factor confirmatory compensatory and the bi-factor confirmatory partially compensatory multidimensional IRT. In addition, these models are proposed in such as way they can be promisingly extended from the unidimensional IRT and the bi-factor model. The Bayesian estimation approach is soundly suggested to be the most dependable estimation procedure available to produce reliable subscores.

There are four objectives for this study – 1.) to propose the bi-factor confirmatory compensatory and the bi-factor confirmatory partially compensatory MIRT models for improving the reliability of subscore estimates; 2.) to evaluate how well the proposed bi-factor confirmatory compensatory and the partially compensatory MIRT models perform in terms of Bayesian parameter estimation accuracies under various parameter conditions; 3.) to compare model-data fit within each of the bi-factor confirmatory MIRT model in terms of Bayesian complexity and fit under various parameter conditions of the hypothesized true model; and 4.) to quantify the estimated subscores from the bi-factor confirmatory compensatory MIRT model and the estimated subscores of the bi-factor confirmatory partially compensatory MIRT model.

In general, how would a bi-factor model be specified in confirmatory compensatory and confirmatory partially compensatory MIRT models and how would the subscores from both models be interpreted were inquired. The primary research questions for a simulation in this study were:

1. How well do the proposed models perform in recovering item and examinee parameters under various simulation conditions?

2. How would the parameters estimated from the bi-factor confirmatory compensatory MIRT model and the parameters estimated from the bi-factor confirmatory partially compensatory MIRT model be different in terms of mode-data fit under various conditions of the hypothesized true model to achieve reliable subscore estimates?

3. How well can the reliability and classification (validity) of the Bayesian subscores be estimated from both models?

1.5 Hypotheses

The bi-factor confirmatory compensatory and partially compensatory MIRT models are straightforwardly specified from the definitions of the multidimensional IRT framework and the bi-factor model. Procedures for subscore estimation using Bayesian approach are described in Chapter 3. Hypotheses to align research questions (1) through (4) are as follow:

1. Each of the proposed models performed well in recovering item and examinee parameters under various simulated conditions from the true models.

2. Subscores estimated from the proposed models can be mutually described based on the model-data fits of the parameters from the bi-factor confirmatory compensatory or from the bi-factor confirmatory partially compensatory MIRT modesl when reliability of the subscores is improved.

3. Subscores estimated from the bi-factor confirmatory compensatory MIRT model with Bayesian estimation approach and/or subscores estimated in the bi-factor confirmatory partially compensatory MIRT model with Bayesian approach have increased reliability when subscales of the specific content domains are uniquely separated from the primary domain.

1.6 Definitions of Variables

The definitions of terms that are used in this study are summarized as following:

Subscores

- Scores that are derived from subtests and make-up part of an overall test score.

Confirmatory Multidimensional Item Response Theory Model

- A model of Item Response Theory (IRT) signifies multiple dimension items and person parameters that have a clear hypothesis for the structure of the item response data[a].

Compensatory Multidimensional Item Response Theory Model

- A model of Item Response Theory (IRT) which signifies a linear combination of multiple dimensions of item and ability parameters. The linear combination is an additive in nature. Thus, an examinee with low ability in one of the dimensions being measured compensates by having higher ability on the other dimensions[a].

Partially Compensatory Multidimensional Item Response Theory Model

- A model of Item Response Theory (IRT) which signifies a nonlinear combination of multiple dimensions of item and ability parameters. The nonlinear combination is a multiplicative in nature. Thus, scoring high on one dimension of ability cannot compensate for scoring low on an other ability[a].

Bi-factor Model

- A model signifies one primary factor or general targeted skill, and allows for one or more orthogonal secondary factors or specific content domains.

Bi-factor Multidimensional Item Response Theory Model

- A model of compensatory or partially compensatory multidimensional IRT which signifies a linear or nonlinear combination of primary and orthogonal specific dimensions of item and ability parameters.

Note. [a]Definition adopted from Reckase (2009).

1.7 Summary and Significance of the Study

The motivation for investigating the efficacy of the studied models in estimating subscores are that the estimated subscores could provide diagnostic or quality information for test users and stakeholders, and thus could be useful for practical purposes. As the subscores estimated from the studied models established reliability and validity for comparability, subscores may be reported cautiously as for diagnostic purposes, teacher accountability evaluations, and curriculum effectiveness evaluations.

Simultaneous consideration of the multiple dimensions of the examinee's abilities and item characteristics should be considered as one of the major concerns when reporting subscores. Thus, the uses of the appropriate sophisticated approaches that used in this study are important to gather as much information as possible in reporting reliable and valid subscores. The sophistication of the models are applied to obtain the most accurate measurement possible of the mental abilities of the examinees, which requires accurate measures of precisions for informed judgment and future remedial actions.

Chapter 2 reviews various existing studies on subscore estimation methods and reliability of the estimated subscores. The item response theory framework and the bi-factor approach for subscore estimations are discussed. Also, challenges from the perspectives of the model construction and application are briefly presented. Considerable works of research and studies that have been done in the multidimensional item response theory and the bi-factor models are highlighted.

Chapter 3 defines the studied models specifications from the bi-factor compensatory and partially compensatory multidimensional IRT. Then, simulation study designations from the studied models are explained. Evaluation criteria that are for the simulation study and evaluation criteria for the scoring procedures based on the studied models are addressed. Chapter 4 demonstrates results from the simulation study and Chapter 5 is devoted for discussion and conclusion.

Chapter 2

Literature Review

This chapter begins with a prevalent technique on improving the precision of subscore estimates based on a regression approach that is known as Kelly's regression for true score estimate (Kelly, 1927, 1947). An extension of Kelly's regression technique is followed by a discussion of the idea of an augmented subscore - that is, it is introduced as a multivariate version of Kelly's regression (Wainer et al., 2000, 2001). Other approaches for subscore estimations to improve reliability and validity are briefly summarized. Then, studies on multidimensional techniques for subscore augmentation and studies on the bi-factor model for general and specific subscore estimations are discussed. Research findings on comparisons of existing subscore estimation techniques are also presented throughout the chapter. Finally, arguments about the multidimensional item response theory (MIRT) with the bi-factor model as another promising approach to estimating reliable subscores are discussed as well as adaption of issues related to subscore estimations that are presented in the literature .

2.1 Regression Approach of Subscore Estimates

Kelly's univariate regression (Kelly, 1927, 1947) can be written as

$$\widehat{\tau} = \rho x + (1 - \rho)\, \mu \tag{2.1.1}$$

in which an estimated true score, $\widehat{\tau}$, is improved by shrinking the observed score, x, toward the group mean, μ, by an equal amount of reliability, ρ. The first part of the Kelly's regressed score in Equation 2.1.1,ρx, tells that the estimated true score retains a reliable score and the second part, $(1-\rho)\mu$ is to remove the unreliable score by regressing it toward the group mean. Thus, this approach improves the reliability of the true score by using the group mean. As the estimated observed score becomes increasingly reliable, the estimates of true score would be very close to the observed score. In contrast, as the estimated observed score becomes less reliable, the estimates of the true score would shrink toward the group mean.

Using the information from the rest of the test to improve the precision of the estimates is analogous to the empirical Bayes estimation (EB) (Wainer et al., 2001; Yen, 1987) based on ancillary information to increase the precision of subscore estimates. When a sample mean estimate, x, and a sample reliability estimate, r, are substitute into Equation 2.1.1 above and rearrange the equation, that is

$$\widehat{\tau} = rx + (1-r)x. = x. + r(x - x.),$$

thus, this Kelly's univariate regression of true score on observed score can be generalized for multivariate cases such that

$$\widehat{\tau} = \mathbf{x}. + \mathbf{B}(\mathbf{x} - \mathbf{x}.), \tag{2.1.2}$$

where $\mathbf{x}.$ is a vector of subscale means, $\mathbf{x}$ is a vector of subscale scores and $\mathbf{B}$ is the matrix of the reliability based weights. Thus, $\widehat{\tau}$ is an improved estimate of the subscale true scores based on the examinee's performance for each subscale and performance on the other subscales. Recall that the index of reliability, ρ, can be defined as the proportion of true score variance relative to the observed score variance and is directly estimable from the data, for example by using the Split Half method and Cronbach's coefficient alpha. Thus, $\mathbf{B}$ can be estimated from the product of the true score covariance matrix to the inverse of the observed score covariance matrix, that is

$$\mathbf{B} = \sum_t \sum_x^{-1}$$

where $\sum_t$ is the true score covariance matrix and $\sum_x^{-1}$is the inverse of the observed score covariance matrix. As shown by Wainer et al. (2001), this is the conventional notation for matrix of regression coefficients, that the weights for a linear combination of the deviance scores are used to best estimate the subscale scores, $\widehat{\tau}$. The multivariate reliability-based weights, **B**, can also be defined as the multivariate index of reliability and thus can be directly estimated from data: $\sum_t$ can be estimated by the covariance matrix of the true score, $\mathbf{S}_\mathrm{t}$, and $\sum_x^{-1}$can be directly estimated by the covariance matrix of the observed score, $\mathbf{S}_\mathrm{x}$. For $\mathbf{S}_\mathrm{t}$, it can be computed from the estimated observed scores covariance matrix, element-wise using

$$S^t_{dd'} = S^x_{dd'} \mathrm{for}\, d \neq d'$$

and

$$S^t_{dd'} = \rho_d S^x_{dd'} \mathrm{for}\, d = d'$$

where $S^t_{dd'}$is the dd'element in the covariance matrix $\mathbf{S}_\mathrm{t}$ and $S^x_{dd'}$ is the dd'element of the covariance matrix $\mathbf{S}_\mathrm{x}$ and ρ_d is the reliability of the subscore d. In addition, it is known that the lower bound of reliability is a customary from the Cronbach's coefficient alpha,α, and is shown as

$$\rho \geq \widehat{\alpha} = \frac{n}{n-1}\left[1 - \frac{\sum_{j=1}^{n} \sigma^2_{y_j}}{\sigma^2_x}\right] \equiv \frac{\sigma^2_t}{\sigma^2_x}, \qquad (2.1.3)$$

where n is the number of items on the subscale. Cronbach's alpha is calculated from the observed item variances, $\sigma^2_{y_j}$, and total test score variance, σ^2_x, and thus it is computational equivalent to the ratio of true score variance to the total score variance (Crocker & Algina, 1986).

Suppose there are two variables, y_1 and y_2 that are multivariately-normally distributed, that is notationally:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

Then, using this standard analogy, assuming that the true score, τ, and the observed score x, follow a multivariate normal distribution with a common mean, μ, that is notationally:

$$\begin{bmatrix} \tau \\ \mathbf{x} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma_t & \Sigma_t \\ \Sigma_t & \Sigma_x \end{bmatrix}\right).$$

Following Equation 2.1.2and substitutes $\mathbf{B} = \sum_t \sum_x^{-1}$into the equation, the empirical Bayes estimate of the vector of true subscale scores for examinee i, τ_i is

$$E\left(\tau_i \mid \mathbf{x}_\mathrm{i}\right) = \mu + \sum_t \left(\sum_x \right)^{-1} \left(\mathbf{x}_\mathrm{i} - \mu\right) . \tag{2.1.4}$$

For perfectly a reliable test, that is when $\rho = 1$, the empirical Bayes estimate of the true score is equal to the observed score. Consequently, for a perfectly reliable test containing subscales, that is when $\mathbf{B} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, the empirical Bayes estimate of the vector of the true subscale scores is equal to the observed subscale scores. When the error of measurement increases, the second term in Equation 2.1.2 shrinks toward zero and thus the best estimate of the true score is the mean over all examinees. By adapting the reliability index in Equation 2.1.3, the estimate of the subscale reliability can be computed as the ratio of the true score variance of the dth subscale to the estimated true score variance of the dth subscale and can be written as

$$\rho_d^2 = \frac{a_{dd}}{c_{dd}} \tag{2.1.5}$$

wherea_{dd} is the true score variance of the dth subscale, which is the diagonal element of the matrix,

$$\mathbf{A} = \mathbf{S}_\mathrm{t}\left(\mathbf{S}_\mathrm{x}\right)^{-\mathbf{1}} S_\mathrm{t}\left(\mathbf{S}_\mathrm{x}\right)^{-1} S_\mathrm{t},$$

and $c_{dd}$is the estimated score variance of the dth subscale, which is the diagonal element of the matrix,

$$\mathbf{C} = \mathbf{S}_{\mathrm{t}}\mathbf{S}_{\mathbf{x}}^{\mathbf{-1}}S_{\mathrm{t}}.$$

The estimate of the reliability in Equation 2.1.5 has shown to be positively biased (Edwards, 2002). Instead, Edwards (2002) and Edwards & Vevea (2006) recommended reliability for the augmented subscore to be computed from the diagonal element for the numerator of

$$\mathbf{A}^{\star} = \mathbf{S}_{\mathrm{t}} - S_{\mathrm{t}}\left(\mathbf{S}_{\mathrm{x}}\right)^{-1}S_{\mathrm{t}},$$

and the diagonal element for the denominator is taken from the true score matrix, thus,

$$\widehat{\rho}_d^2 = 1 - \frac{a_{dd}^{\star}}{s_{dd}^{t}}. \tag{2.1.6}$$

2.2 Subscore Augmentation from IRT Scale Score

Many large-scale testing programs used the IRT scale score for viable scoring and score reporting, as the IRT scale scores can be computed from items that have more than one category and the scores are comparable across test forms. As the IRT models provide useful properties for subscore augmentation, Wainer et al. (2001) proposed that the same technology of the empirical Bayes estimate be applied to the IRT scale score estimate of latent ability,θ, like the maximum a posteriori estimate $(\mathrm{MAP}[\theta])$ or expected a posterior estimate $(\mathrm{EAP}[\theta])$.

The value of $\mathrm{MAP}[\theta]$ and $\mathrm{EAP}[\theta]$can be computed for each examinee for each subscale. $\mathrm{MAP}[\theta]$ and $\mathrm{EAP}[\theta]$ are two estimates that are analogous to the Kelly's regressed estimates. As the Kelly's regressed estimate shrinks linearly or proportionally toward the group mean, $\mathrm{MAP}[\theta]$ and $\mathrm{EAP}[\theta]$ estimates shrink toward the mean proportional to their variance. Therefore, $\mathrm{MAP}[\theta]$ and $\mathrm{EAP}[\theta]$ estimates shrink more toward the mean as the response patterns provide less informative and $\text{MAP}[\theta]$ and $\text{EAP}[\theta]$ estimates shrink less when there is more information in the response patterns.

Refer $\text{EAP}[\theta_d]$ (or$\text{MAP}[\theta_d]$) to the IRT scale score of response-pattern or summed-score for subscale d. Assuming that the mean subscore for each subscale is zero. For a well-constructed test, the variability in the precisions of the linear combination of $\text{EAP}[\theta]$s (or $\text{MAP}[\theta]$s) is relatively very small and thus can be ignored (Wainer et al., 2001). Therefore, for subscore augmentation from these IRT scale score estimates, an $\text{EAP}[\theta]$ (or $\text{MAP}[\theta]$) can be treated as observed scores. Then, by the analogy with Kelly's regression

$$\text{EAP}[\theta_d] \approx \rho_d \text{EAP}^{\star}[\theta_d] \tag{2.2.1}$$

where ρ_d is an estimate of reliability of subscale d, $\text{EAP}^{\star}[\theta_d]$ is the hypothetical IRT scale estimate from the EAP estimate of the latent variable on subscale θ_d. Note that, $\text{EAP}^{\star}[\theta_d]$ is not regressed toward the mean but the value of $\text{EAP}^{\star}[\theta_d]$ is identical to the observed summed score for subscale d. Solving Equation 2.2.1

$$\text{EAP}^{\star}[\theta_d] \approx \frac{\text{EAP}[\theta_d]}{\rho_d} \tag{2.2.2}$$

The denominator of Equation 2.2.2) is the marginal reliability, and as proposed by Green et al. (1984). This marginal reliability can be estimated from

$$\widehat{\rho}_d = 1 - \bar{\sigma}_e^2. \tag{2.2.3}$$

$\bar{\sigma}_e^2$ in the above equation can be computed from

$$\bar{\sigma}_e^2 = \int E\left(\sigma_e^2\right) \phi(\theta)\, d\theta,$$

where $E\left(\sigma_e^2\right) = \sum_{all\,x} \sigma_e^2 L(x \mid \theta)$ and x is the summed score, and $L(x \mid \theta)$ is the joint likelihood for summed score x from the item response theory model. The $\text{EAP}^{\star}[\theta_d]$ for each examinee for subscale d is then computed by substituting $\widehat{\rho}_d$ into Equation 2.2.2 above.

Similar to the observed score variance-covariance matrix,$\mathbf{S}_\mathrm{x}$, of the Kelly's regression technique, the element in the variance-covariance matrix of that observed $\mathrm{EAP}^\star[\theta_d]$, that is notated as $\mathbf{S}_\mathrm{EAP}$, among the original IRT scale scores is corrected and can be computed from

$$S^x_{dd'} = \frac{S^\mathrm{EAP}_{dd'}}{\rho_d \rho_{d'}} \,\mathrm{for}\, d \neq d'$$

and

$$S^x_{dd} = \frac{S^\mathrm{EAP}_{dd'}}{\rho_d^2} \,\mathrm{for}\, d = d'$$

Thus, for examinee i, the empirical Bayes estimates of the IRT scale scores can be computed from Equation 2.2.4 below. This is analogous to Equation 2.1.4 presented above.

$$\underline{\mathbf{E\hat{A}P}}[\theta] = \underline{\mathbf{EAP}}[\theta]\,. + \mathbf{B}^\star\left(\underline{\mathbf{EAP}}^\star[\theta]_i - \underline{\mathbf{EAP}}[\theta]\,.\right) \qquad (2.2.4)$$

where for examinee i, $\mathbf{EAP}^\star[\theta]_i$is the vector of IRT scale scores for the subscale, $\underline{\mathbf{EAP}}[\theta]\,.$is the average vector of the IRT subscale scores, $\underline{\mathbf{E\hat{A}P}}[\theta]$is the vector of empirical Bayes estimate of the IRT scale scores and $\mathbf{B}^\star = \mathbf{S}^t\left(\mathbf{S}^x\right)^{-1}$. To simplify, the diagonal elements in the vector of reliability-based weights can be computed as

$$b_{dd'} = \frac{b^\star}{\rho_{d'}},$$

that the weights $b_{dd'}$ for subscale d is regressed on subscale d' in $\mathbf{B}^\star$. The augmented estimates from the original values of $\underline{\mathbf{EAP}}[\theta]$ can be computed from the weights $\mathbf{B}$.

Another approach to subscore augmentation is similar to the Wainer et al. (2001) by means of using an empirical Bayes procedure that is proposed by Yen (1987). The procedure is to pool performance on particular items representing some specific test objective from the overall examinee's performance. This procedure is a basis of the objective performance index (OPI) and is used for some tests published by CTB/McGraw-Hill. The OPI approach is closely related to Wainer et al. (2001) discussed previously. Wainer et al.'s and Yen's methods for the subscore augmentation are different that Yen's procedure which is based on the binomial distribution of the scores whereas Wainer et al.'s is based on the normal distribution that forms the basis of the Kelly's regressed true score estimates (Kelly, 1927, 1947).

Unlike Yen's OPI procedure, the Wainer et al.'s subscore augmentation can be extended for multivariate cases and, thus, it is well suited for a test that is truly multidimensional. In other contexts, such as to improve subscore estimations, is by means of using the collateral or auxiliary information in the IRT about the examinee. Collateral information is adaptively measured in a test. The information is an item provided to explain secondary ability of separate content areas (Ackerman & Davey, 1991). According to Stout et al. (2003), collateral information refers to additional estimation information derived from variables that are distinct from, but correlated with, the studied relevant variable of interest. Collateral information describes a dominant activity or process and is useful for item calibration that could provide practical important gains in item calibration accuracy.

Auxiliary information sometimes referred to as ancillary information, can be any related factors on the test that are correlated with the latent proficiency such as age, courses taken, years of schooling, grade obtained and even family background (Mislevy, 1987; Wainer et al., 2000; Wang et al., 2004). Collateral information has not been adopted sensibly from the contexts presented by Ackerman & Davey (1991) and Stout et al., (2003). However, augmentation with auxiliary information is treated as "collateral information" in many recent studies to improve subscore estimates (e.g. Tao (2009)). As there have been so many broad perspectives of collateral (or auxiliary) information adopted, subscore augmentation can be a special case of using collateral (or auxiliary) information to improve score precision. Therefore, collateral or auxiliary information can be any of the factors that are correlated with the objectives of a test.

For a firm definition, de la Torre & Hong (2010) described these two terms as out-of-test collateral information and in-test collateral information. The out-of-test collateral information is similarly defined from the context of Mislevy (1987) and Wang et al. (2004) as any demographic educational variable. For example, NAEP uses the out-of-test collateral information from student and school characteristics to obtain scores that are more accurate for various subpopulations of students. The In-test collateral information is analogous to the context of ability estimation in the IRT framework by Ackerman & Davey (1991) and Stout et al. (2003) which is inherent information that can be found in the examinees' item responses, specifically, responses to other test domains. Studies have shown that, for the IRT ability estimation, incorporating in-test collateral information in the scoring process has improved the precision of ability estimates (Wainer et al., 2001; Yen, 1984; de la Torre & Patz, 2005; de la Torre & Song, 2009; Wang & Gao, 2010).

2.3 Mutidimensional IRT Subscore Estimation

Much attention has been given to MIRT models to their abilities to improve subscore estimates, and the consequences for scoring and score reporting. Due to several variants of multidimensional item response theory (MIRT) models, this is a lack of solid operational research on these variants and there are limited options for IRT software for carrying out item calibration and parameter estimations. MIRT models allow high ability in one dimension to compensate for low ability on other dimensions, which is much more common, and such a model is called the compensatory MIRT model. For the noncompensatory model (no such compensation), high probability means high ability for all dimensions (Reckase, 1997, 2009).

Application of MIRT models is also being considered in a great deal for exploratory and confirmatory types of analysis. According to Reckase (2009), the MIRT is still in its developmental infancy, and making important advances in the procedures could lead to a better understanding of test functions in multidimensional space of ability and cognitive proficiencies. Thissen & Edwards (2005) introduced a simultaneous estimation procedure with a potential solution using the Markov Chain Monte Carlo methods (MCMC) for the constrained MIRT model. MCMC estimation for IRT (2PL and 3PL) was proposed by Patz & Junker (1999a) and Patz & Junker (1999b) without using data augmentation as the studies used Metropolis-Hasting with Gibbs sampling. Thissen & Edwards (2005) and Edwards (2010)applied a data augmentation strategy that relied upon the Gibbs sampling (Geman & Geman, 1984) approach for the three-parameter normal ogive model and Samejima (1969) graded response model. The MCMC method produced efficient estimates of subscores for sets of items that are or are not mutually exclusive. Although the studies observed the MCMC simulation for more complex model is a bit longer than accustomed waiting time for most researchers, the MCMC simulation provides the entire posterior distribution of every parameter, and not just a point estimate and standard error, which can then be used for further evaluation of the parameter estimates.

A study by de la Torre & Patz (2005) which was based on the 3PL-MIRT model of Reckase (1997) and Reckase (2009) and estimated within a Metropolis-Hastings of the MCMC algorithm (Chib & Greenberg, 1995; Gilks et al., 1996) which is closely related to empirical Bayes approach Wainer et al. (2001). The study focused on correlated ability, and multidimensional scores could be used to inform finer-grained reporting such as skills profiles and objective level scores. The study defined a multidimensional approach to simultaneous ability estimation that can be viewed as a more general framework and straightforward for obtaining expected a posteriori estimates of ability. Note that the study applied the 3PL-MIRT model from Reckase (1997) and Reckase (2009) and assumed hierarchical ability estimates. Thus, there is no assumption about the compensatory or noncompensatory approach.

From the same study, ability from the hierarchical model can be estimated from the multidimensional expected a posteriori (EAP-M), and all together can be computed by adding up EAPs across the number of dimensions of ability over thousands of the MCMC iterations. Pearson correlation and mean squared error (MSE) were used to evaluate the correspondence between the estimated and the generated abilities. Unlike Wainer et al. (2001), de la Torre & Patz (2005) did not use the marginal reliability in Equation 2.2.3 above. Instead, reliability of subscale d is defined to analogous Wainer's et al. (2001) "MAP$[\theta]$" approach of response pattern estimates of subscore, that is

$$\hat{\rho}_d = \frac{var\left(EAP\left[\theta_d\right]\right)}{var\left(EAP\left[\theta_d\right]\right) + \overline{Pvar\left(EAP\left[\theta_d\right]\right)}}$$

where $\overline{Pvar\left(EAP\left[\theta_d\right]\right)}$is the average posterior variance of ability estimates in subscale d. The study defined the corrected covariance matrix for reliability as

$$\mathbf{S}^c = \mathbf{S}^u - \mathbf{D}$$

where $\mathbf{S}^u$ is the covariance matrix of the unregressed ability estimate $\text{EAP}^{\star}\left[\theta_d\right]$ (as in the Equation 2.2.2 above) and is analogous to $\mathbf{S}^x$ (i.e. the observed score variance-covariance matrix of Kelly's regression technique), and $\mathbf{D}$ is the diagonal matrix whose dth nonzero entry is $\left(1-\hat{\rho}_d\right)\mathbf{S}^u_{dd}$. Thus, the empirical Bayes estimate for examinee i is given by

$$\underline{\mathbf{EAP}}\left[\theta^{(1)}\right]_i = \underline{\mathbf{EAP}}\left[\theta\right]. + \mathbf{S}^c\left(\mathbf{S}^u\right)^{-1}\left(\underline{\mathbf{EAP}}^{\star}\left[\theta\right]_i - \underline{\mathbf{EAP}}\left[\theta\right].\right). \tag{2.3.1}$$

The study compared the empirical Bayes estimates, $\theta^{(1)}$, in Equation 2.3.1 above with the ability estimates obtained using simultaneous estimation, $\theta^{(0)}$, from the EAP-M method explained in the paper. The correlation between EAP-M of the simultaneous estimation via MCMC is almost perfect, even with the worst simulation conditions, with the estimate from EAP the Equation 2.3.1 above. The difference in the MSE between the two methods is very small up to three decimal places. They observed that $\theta^{(0)}$ is slightly better than $\theta^{(1)}$as the simultaneous estimation of subscores via MCMC produced smaller error and standard error than the approach taken by Wainer et al. (2001).

Estimating subscores from the multidimensional IRT is not always straightforward. Subscores can be estimated from the three-parameter logistic MIRT model with oblique simple-structure (Thurstone, 1947) for a common factor model and the bi-factor model (Gibbons & Hedeker, 1992). This study also demonstrated important features of graphical representations of the subscore to provide multivariate profile plots and competency areas from the estimated subscores that may be useful for examinee's diagnostics feedback.

2.4 Bi-factor Model as a Confirmatory MIRT Model

The idealization of the MIRT model to fitting in multiple dimensions of examinee proficiencies and items responses is related to testing conditions in which the test is not strictly unidimensional due to various item contents, complex underlying constructs, and different item types. The MIRT model is an extension model of factor analysis that occupies two major schools of thoughts – the Spearman's two-factor theory and the Thurstone's simple structure of multiple factors of primary abilities. Modernized MIRT models in psychometric and psychology framework designed a plausible model to exhibit a general factor and multiple specific factors. The bi-factor factor analysis(Holzinger & Swineford, 1937) which is the special case of hierarchical and higher order factor analysis is merely an extension of Spearman's two-factor model. The bi-factor model for mental tests and psychological measures conceived item loadings on the general factor plus additional specific domain factors.

The term "bi-factor" largely appears in confirmatory factor analysis (Jöreskog, 1969), but it is straightforward to conceive at the item-level pattern (Muthén, 1989). McLeod et al. (2001) and Reckase (2009) translated the associated factor analysis parameters into their multidimensional IRT analogs. Pairing the bi-factor model with the IRT model was first illustrated in 1990 for selected items from an ACT natural science test in Gibbons et al. (1990). The technical report then appeared in the 1992 Psychometrika journal extending the bi-factor IRT model to the psychiatric symptom rating scale (Gibbons & Hedeker, 1992). The focus of the research was on binary items in the unidimensional ability continuum. For constructed response items, Gibbons and colleagues actively developed the bi-factor unidimensional IRT model a decade and a half later for Samejima's graded response model (e.g. see Wainer et al., 2001).

The bi-factor model represents all underlying items to have positive loadings, which are similar to the positive regression coefficient on a general trait. Typically the general trait will be conceptually broader and is used in scaling individual proficiency, process, or target skill. In addition, an item can load on zero, one, or more within a cluster of items that measure specific content domains or "group" or "method" related factors. These factors are assumed to be orthogonal to each other

(Gibbons & Hedeker, 1992; Gibbons et al., 1990; Reise et al., 2007, 2010; Cai, 2010).

A series of studies in Reise et al. (2007), Reise et al. (2010) and Reise et al. (2011) applied the bi-factor model to the Consumer Assessment of Healthcare Providers and Systems (CAHPS©2.0) survey instrument and personality assessments. The studies focused on the examination of the distortion of the item discrimination parameter due to local dependencies caused by the secondary influences. The study also suggested the bi-factor model as evidence in forming subscales as each item set was an adequate indicator of its respective construct, albeit the correlation between dimensions was moderate to large ($r = .60$) in the exploratory factor analysis, and $r = .74$ in the confirmatory analysis). The studies concluded from the bi-factor analysis that the results showed clearly that each dimension was, in fact, confounded by both general and specific sources of variance that might not be detected in the MIRT solutions.

The MIRT bi-factor model has received less attention in the literature. One study demonstrated a simulation study on 2,000 examinees as well as an empirical example on 5,000 examinees for math and reading tests from the Programme for International Student Assessment 2000 (PISA 2000) by DeMars (2006). The study compared the multidimensional bi-factor IRT model with testlet-based models, where the application of testlet-based models can be found in Li et al. (2006) and Bradlow et al. (1999) and testlets-as-polytomous-item model in Lee et al. (2001) and Sireci et al. (1991). Those studies focused on the investigation of the effects of number of items for each dimension, number of dimensions, and correlations between dimensions on ability estimates.

Gibbons & Hedeker (1992), Gibbons et al. (2007) and Gibbons et al. (2009) examined instruments in the Psychiatric Diagnostic Screening Questionnaire (PDSQ) and Post-Traumatic Growth Inventory (PTGI). The studies investigated that the multidimensional bi-factor IRT model provides more homogenous results and more reliable estimates of the underlying dimensions than the unidimensional Samejima's graded response model. Also, Immekus & Imbrie (2008) observed that modeling the bi-factor model within an IRT framework is feasible for a confirmatory-based method to collect construct validity evidence of obtained scores.

A Bayesian revival, particularly the Markov Chain Monte Carlo methods demonstrated in de la

Torre & Song (2009), Edwards (2010), Shin (2007), Stone et al. (2010), Tao (2009), Tate (2004), Yao (2010), and Sheng & Wikle (2009) showed improvement in parameter estimations as well as subscore estimates. Edwards (2010) revealed that posterior distributions of ability estimates from the multidimensional bi-factor model provide the flexibility to add additional features and complexities of the multidimensional test structure for study. In the study, the author suggests more research is needed to determine the effect of the strength of prior distributions for items and ability parameters besides the models that are problematic with the current "gold standard" estimators. Unlike MCMC, a non-Bayes estimator such as full-information maximum likelihood with EM algorithm estimator (FIML-EM), for example in Gibbons & Hedeker (1992), Gibbons et al. (2007), Immekus & Imbrie (2008) and Rijmen (2010), is problematic in that the method performed inadequately when confronted with problems requiring high dimensional numerical integration and multiplication of the probability matrices in the bi-factor model. Thus, the Bayesian approach is a promising estimation alternative method for the bi-factor model within a multidimensional item response theory framework.

Chapter 3

Methodology

This chapter is organized with four main sections. The first section discusses the specification of the confirmatory compensatory and confirmatory partially compensatory multidimensional IRT (MIRT) models, in particular, the 3PL multidimensional IRT models. The models are analogous to the bi-factor model to estimate subscores for primary dimension and specific ability dimensions, will later be called bi-factor compensatory (BF-C) and bi-factor partially compensatory (BF-PC) MIRT models. Bayesian approach to estimate subscores from BF-C and BF-PC is also addressed. The second section explains the design for the Monte Carlo simulation study including dependent and independent variables for this study. The following section discusses criteria for evaluation of the accuracies of the parameter estimations from both the bi-factor compensatory and the bi-factor partially compensatory MIRT models. Finally, procedures for evaluating the estimated subscores are proposed including reliability and classification of the subscores.

3.1 Models for Subscores Estimation

3.1.1 Confirmatory 3PL MIRT Model

Test total score and subscores for test with multidimensional structure can be estimated from the MIRT model. One of the MIRT models for multiple-choice or binary items is the multidimensional extension of the three-parameter logistic IRT model (Birnbaum, 1968) or the MIRT 3PL model by Reckase (1985, 1997 & 2009). From this model, test scores for an examinee can be estimated from a combination of multiple dimensions of examinee abilities and item characteristics.

In general, there are two types of MIRT models. One is from a linear combination of multiple dimensions of abilities. This model is known as the compensatory MIRT model. The model explains the compensatory nature of examinee abilities to response on an item that is a high ability on one dimension can compensate for a low ability on another dimension to yield probability of responses as more moderate and balance values. Thus, low performance in one dimension is additive with the other higher performances, and that gives the same overall performance. The logistic function of the compensatory M3PL model can be written as

$$P\left(U_{ij}=1\mid\xi\right)=c_j+\left(1-c_j\right)\frac{exp\left\{\sum_{d=1}^{D}a_{jd}\left(\theta_{id}-b_j\right)\right\}}{1+exp\left\{\sum_{d=1}^{D}a_{jd}\left(\theta_{id}-b_j\right)\right\}} \tag{3.1.1}$$

and can be rewritten as a probit function as

$$P\left(U_{ij}=1\mid\xi\right)=c_j+\left(1-c_j\right)\Theta\left(\sum_{d=1}^{D}a_{jd}\left(\theta_{id}-b_j\right)\right) \tag{3.1.2}$$

which is also known as the normal ogive or the standard normal model that is denoted by Θ symbol.

In Equation 3.1.2, $P\left(U_{ij}=1\mid\xi\right)$is the probability for an examinee i answering jth item correctly $\left(U_{ij}=1\right)$with respect to item and ability parameters denoted as $\xi=(\theta,\mathbf{a},b,c)$ where

θ_{id} is the ith examinee's ability for dth dimension for $d=1,2,\ldots,D$ and there are a total of D dimensions of abilities being measured in a test. a_{jd} is the jth item discrimination on dth dimenson and b_j and c_j is the item difficulty and pseudo-guessing parameter, respectively.

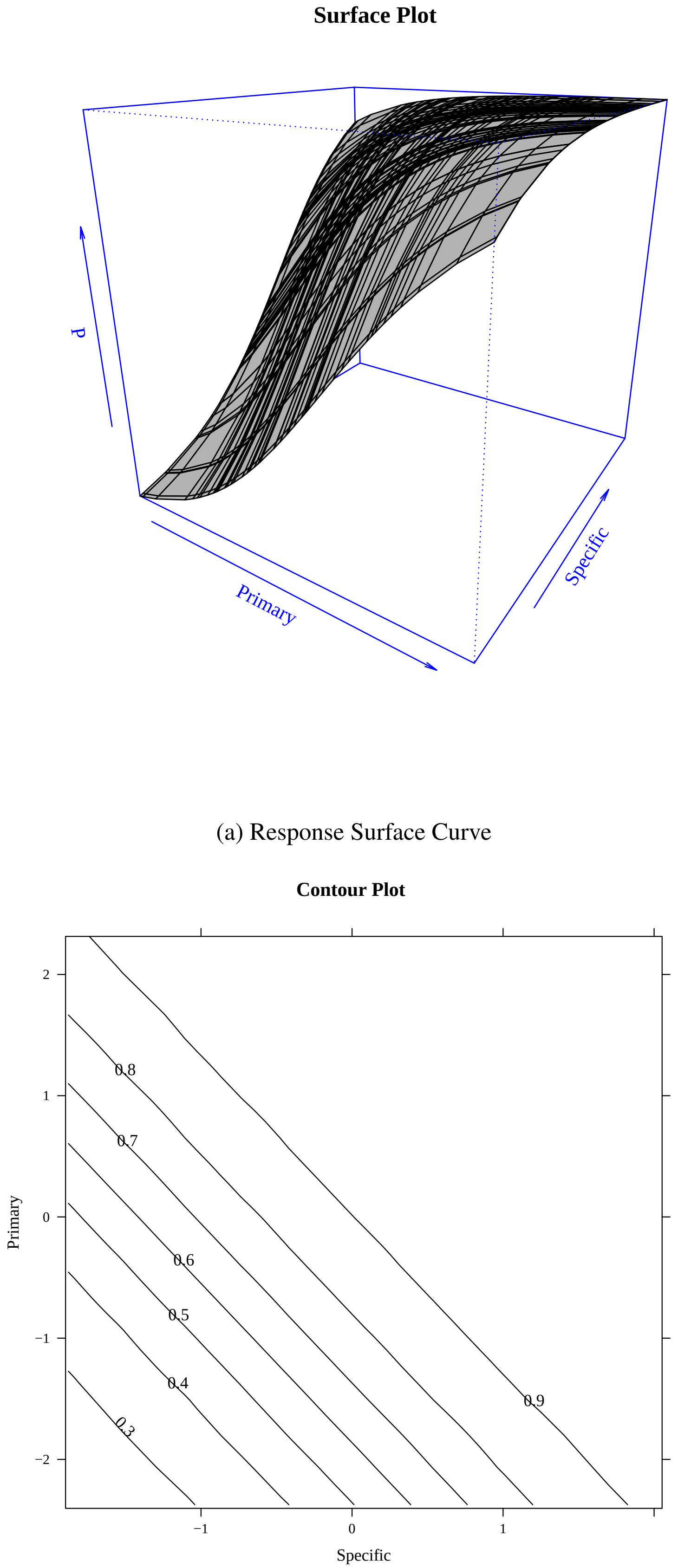
$\theta_{i1},\theta_{i2},\ldots,\theta_{iD}$ are independent identically distributed (iid) random abilities sample from multivariate normal distribution, $\theta\sim MVN(0,\Sigma)$ with a vector of zero means and variance-covariance matrix Σ.

$\sum_{d=1}^{D} a_{jd}\left(\theta_{jd} - b_j\right)$ is a function for a linear relationship between item and ability parameters. Thus, the exponent of the function shows a summation of multiple dimensions of abilities on an item j.

In the compensatory MIRT model in Equation 3.1.1 or Equation 3.1.2, being high on one ability can compensate a low ability on other ability. Figure 3.1 illustrates a compensatory MIRT model. The surface plot for this model shows the probability increases quickly as primary ability increases than it does as specific ability increases. This is because of the different sizes of the a-parameters in the BF-C model. The contour plot posits the location of an examinee in the ability space and the probability of a correct response to an item. This plot also illustrates the compensation between low and high ability. For example, an examinee with low ability on one dimension (e.g. specific content ability on the x-axis), say at -1.0, and with a higher ability on the other dimension (e.g. primary ability on the y-axis), say 1.5, thus the compensation of these abilities makes the overall ability stays on the coordinate (-1.0, 1.5) which is falls onto the .90 line of the contour plot. This examine then has about 90% of the chance to answer an item correctly.

The second type of the MIRT model is the multiplicative model. For three-parameter logistic model, this model is typically called the partially compensatory M3PL model, and also known as the noncompensatory M3PL model. In the partially compensatory MIRT model, being high on one ability cannot compensate for being low in the other dimension of ability. This implies that high performance in one dimension not necessarily but partially yields a higher probability of correct response for the low performance. The linear function in the probability of correct response in Equation 3.1.1 and Equation 3.1.2 are now the product of the probabilities from each dimension. Consecutively, the overall performance is bounded by the lowest probability of success. The mathematical expression of the partial compensatory model can be written as

$$P\left(U_{ij} = 1 \mid \xi\right) = c_j + \left(1 - c_j\right) \prod_{d=1}^{D} \frac{exp\left\{a_{jd}\left(\theta_{id} - b_{jd}\right)\right\}}{1 + exp\left\{a_{jd}\left(\theta_{id} - b_{jd}\right)\right\}} \tag{3.1.3}$$



(a) Response Surface Curve

(b) Item Probability Contours

Figure 3.1: Plots of the Bi-factor Compensatory Model

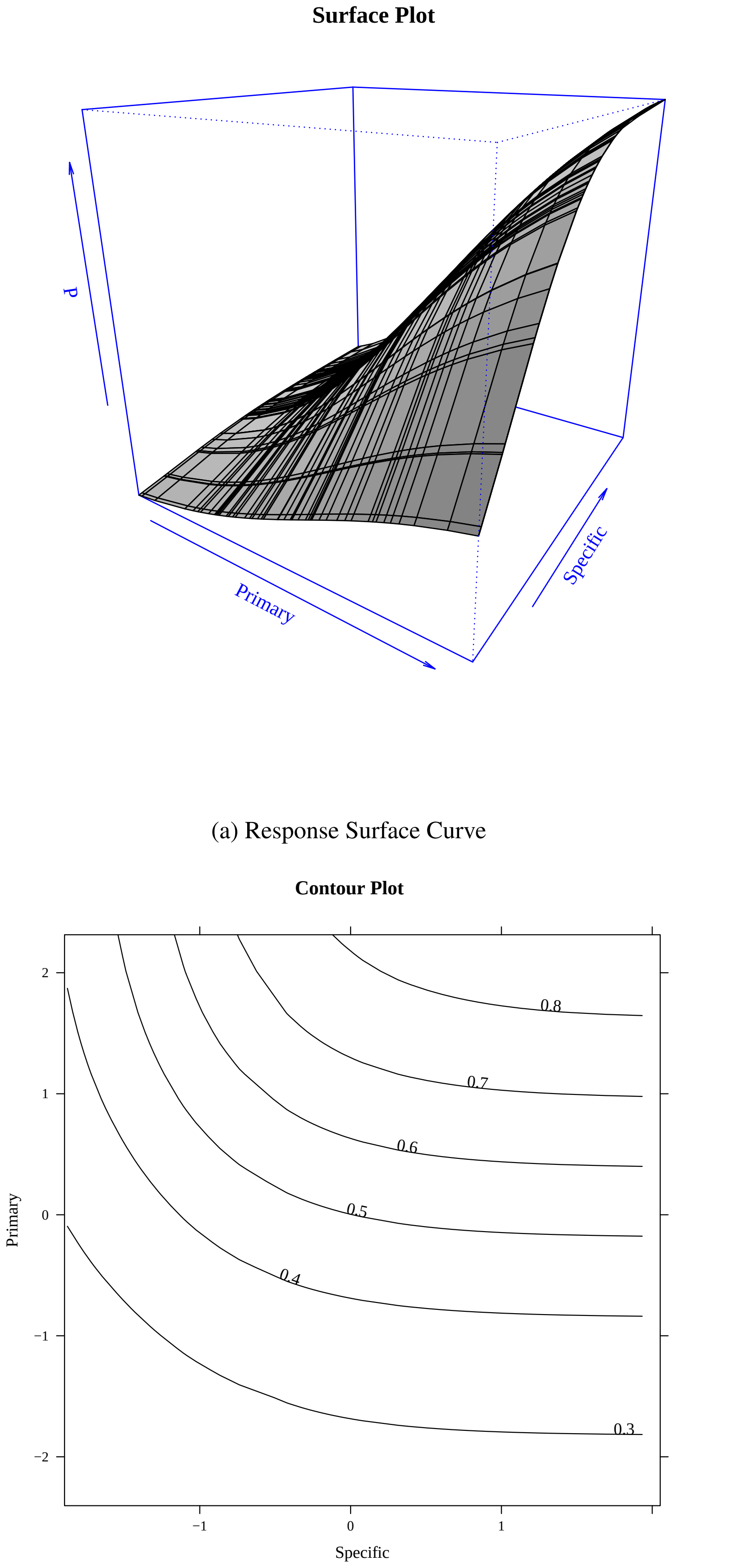
and can be simplified as the probit link function as

$$P\left(U_{ij}=1 \mid \xi\right)=c_j+\left(1-c_j\right)\prod_{d=1}^{D}\Theta\left[a_{jd}\left(\theta_{jd}-b_{jd}\right)\right] \tag{3.1.4}$$

where all the symbols and notations have the same meanings as the previous model, but for partial compensatory model the difficulty parameter for item j is corresponded to the content domain being measured, d, that is denoted as b_{jd} . The product of probabilities from multiple dimensions of abilities reflects the independent of activities by an examinee in response correctly for each dimension. Also, as the number of dimensions increases the probability of success increases.

Figure 3.2 illustrates a partially compensatory MIRT model. The surface plot for this model shows the probability increases slowly as primary ability increases as well as the specific ability increases. This is because of the different sizes of the a-parameters and nonlinear relationship explained be the BF-PC model. The contour plot posits the location of an examinee in the ability space and the probability of a correct response to an item. This plot also illustrates the compensation between low and high ability. From the contour plot, if an examinee ability is at coordinate (2.0, .05), that is the ability in primary dimension is 1.5 level higher than the specific ability dimension, then the examinee will likely have a little more than 55% of the chance to get an item correct. This examinee needs to have higher ability levels on both dimensions to better the chance in getting the item correct. If only one dimension of ability is substantially better than the other dimensions, then the integration between the two abilities in getting an item correct is gradually increase.

Large-scale tests with underlying multidimensional content domains or standards, for example observed that the 8th grade Mathematics test of the Delaware Student Testing Program (DSTP) which includes numeric reasoning, algebraic reasoning, geometric reasoning, and quantitative reasoning to have four dimensions and one dominant factor. Test with this example underlies on multidimensional structure and thus can be modeled with a confirmatory MIRT model by its clearly



(a) Response Surface Curve

(b) Item Probability Contours

Figure 3.2: Plots of the Bi-factor Partially Compensatory Model

defined vector of discrimination parameter, a_j, to separate items measure a common content domain. The predefined vector of the discrimination parameter depicts a dimension or a specific area of content domain in a test that best measured by a cluster of test items. For each item in the cluster, the a-parameter, a_j, in one dimension of content domain is constrained to be nonzero and zero for all others. For this model, it is assumed that the abilities for an examinee between the administered content domains are uncorrelated.

3.1.2 Bi-factor Model

The feature of a confirmatory MIRT model is analogous to the bi-factor model. There is one primary factor and one or more of the secondary factors that is referred to as the specific content domains. Each item will have a nonzero value for the discrimination parameter, $a_{j\in(d-1)} \neq 0$, corresponding to only one of the specific content domains, and other a-parameters are fixed to zero, $a_{j\notin(d-1)} = 0$. There are a total of $D-1$ dimensions of the specific content domains that best measure the specific content abilities. As in the confirmatory MIRT model, the specific content domains are orthogonal to each other as well to the primary domain. Which implies the covariances (or correlations) among the specific content domains are zero. Therefore, for identification of the factor structure, the means and variances of the primary and specific dimensions be set to zero and one, respectively. In addition, all items within a test are targeted to describe a process skill as a general ability that it is referred as the primary ability being assessed. Here, $a_{j1} \neq 0$. Items within a test which capture a similar general skill or ability are considered to have a common variance on each that is explained by the primary domain.

Let say there is a 10-items test with three specific content domains (e.g. algebra, geometry and calculus) and one primary domain (e.g. arithmetic). In bi-factor model, the position of the a-parameters for the primary and specific dimensions can be illustrated in a simple structure matrix, S, as following

$$S = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ a_{31} & a_{32} & 0 & 0 \\ a_{41} & 0 & a_{43} & 0 \\ a_{51} & 0 & a_{53} & 0 \\ a_{61} & 0 & a_{63} & 0 \\ a_{71} & 0 & 0 & a_{74} \\ a_{81} & 0 & 0 & a_{84} \\ a_{91} & 0 & 0 & a_{94} \\ a_{10,1} & 0 & 0 & a_{10,4} \end{bmatrix}$$

In this structure matrix the primary domain items will have a nonzero value of the discrimination, $a_{j1} \neq 0$, and clusters of items that belong to the defined specific ability dimension have nonzero value of the item discrimination, e.g. $(a_{21}, a_{22}, a_{32}) \neq 0$. This implies all other item discriminations are zeros. In general, a test with n items has clearly defined $D-1$ orthogonal dimensions of specific content domains, one dimension exhibits a targeted primary ability, and that primary ability deliberates from all test items. In overall, there are D dimensions of abilities being assessed in a test.

3.2 Scoring Procedure

The proposed models, the bi-factor confirmatory MIRT models – compensatory (BF-C) and partially compensatory (BF-PC), show that the test score derives from the primary content domain can be reported as a complement to the subscores from the specific content domains. Therefore, more than one score are available for reporting and might be useful in providing feedback about examinee's strengths and weaknesses. The subscore on the primary content domain represents examinee's performance on the targeted skill and each subscore portrays examinee's proficiency in each specific content domain. Both the primary and specific content domains subscores are simultaneously estimated when items are calibrate from the proposed models, linearly from the compensatory model or nonlinearly from the partially compensatory model.

The bi-factor confirmatory MIRT modeling from BF-C and BF-PC for estimating subscores can be the alternative procedure to the subscore augmentation described in Wainer et al. (2001) and the objective performance index (OPI) procedure described in Yen (1987) as well as to the traditional summed-score (or number-correct scale), the unidimensional IRT or the MIRT scoring approach. It is reasonable that different levels of the discrimination parameters between the primary dimension and specific dimensions are to weight subscores in an integrated way. Different levels of item discrimination may be used to explain the accuracy of the estimated subscores or to answer a question on how much separations are in the specific content domains subscores after controlling for the score from the primary dimension. Therefore, the precisions of the estimated subscores are observed and it is emphasized in the Subscore Reliability sections and be evaluated in the results sections in Chapter 4.

3.2.1 Bayesian Estimation Method

In this study, only dichotomously (0/1) scored items were examined. Let U be a matrix of item responses, that is the response pattern matrix for examinee i on item j, $\mathrm{U} = \left[u_{ij}\right]_{N\times n}$, for N number of examinees and n number of items. Thus, the observed response for an examinee on jth item can be defined as

$$u_{ij} = \begin{cases} 0 \quad if \text{ response is incorrect} \\ \\ 1 \quad if \text{ response is correct} \end{cases} \qquad i = 1,2,3,\ldots,N;\ j = 1,2,3,\ldots,n$$

An item is best measures either one of the specific content domains and all items best measure the primary targeted ability. Recall that the probit link functions from Equation 3.1.2 and Equation 3.1.4 are associated with the probability of correct response for the compensatory and partially compensatory model, respectively. Hence that the probability of correctly answer a test item is determine by the primary and specific abilities, additively or multiplicatively. For both models, assuming conditional local independence of test items on item and person parameters, the conditional joint probability function of item responses, U, or the likelihood surface of the function can be written as

$$L(\mathrm{U} \mid \theta) = P(\mathrm{U} \mid \theta) = \prod_{i=1}^{N} p_{ij}^{u_{ij}} (1-p)_{ij}^{1-u_{ij}} \tag{3.2.1}$$

where p is the probability of success and $(1-p)$ is the probability of failure. For z_{ij} that is randomly drawn from a uniform distribution with a minimum value of zero and a maximum value of 1, $z_{ij} \sim Uniform(0,1)$, and let u_{ij} be defined as Bernoulli trial with probability of success, p_{ij} given by Equation 3.1.1 to Equation 3.1.4. Thus, the binary item responses by examinee i on the jth item can be

$$u_{ij} = \begin{cases} 0 \quad if\, z_{ij} \geq \quad p_{ij} \\ 1 \quad if\, z_{ij} < \quad p_{ij} \end{cases} \qquad i = 1,2,3,\ldots,N;\ j = 1,2,3,\ldots,n.$$

Know that θ is assumed to be multivariate normally distribute with a zero means vector and an identity matrix of the variance-covariance matrix, that is $\theta \sim MVN(0,\mathrm{I})$. The diagonal elements of the variance-covariance matrix I of θ are fixed to 1.0 and the off-diagonal elements are zeros, and thus the factor structure from the confirmatory MIRT models in Equation 3.1.1 to Equation 3.1.4 is identified. Let $\xi_{jd} = \left(a_{jd}, b_j, c_j\right)'$as the elements of a vector of item parameters for the jth item in dimension d of the BF-C model in Equation 3.1.1 and Equation 3.1.2, and $\zeta_{jd} = \left(a_{jd}, b_{jd}, c_j\right)'$ as the elements of a vector of item parameters for the jth item in dimension d of the BF-PC model in Equation 3.1.3 and Equation 3.1.4. From Bayes' theorem, the Bayesian estimator for primary and specific abilities is not only characterized by the joint likelihood of Equation 3.2.1. With prior distributions for item and examinee's parameters, the joint posterior distributions for (θ, ξ) and (θ, ζ) in the compensatory and partially compensatory models, respectively, can be expressed as

$$p(\theta,\xi \mid \mathrm{U}) \propto L(\mathrm{U} \mid \theta,\xi)\, p(\theta,\xi) \tag{3.2.2}$$

and

$$p(\theta,\zeta \mid \mathrm{U}) \propto L(\mathrm{U} \mid \theta,\zeta)\, p(\theta,\zeta) \tag{3.2.3}$$

where $p(\theta,\xi \mid \mathrm{U})$ and $p(\theta,\zeta \mid \mathrm{U})$ represent the posterior distribution for the compensatory and partially compensatory model, respectively. $L(\mathrm{U} \mid \theta,\xi)$ and $L(\mathrm{U} \mid \theta,\zeta)$ are the likelihood functions and $p(\theta,\xi)$ and $p(\theta,\zeta)$ are the prior distributions of item and examinee parameters.

3.2.2 Markov Chain Monte Carlo with Gibbs Sampling

Subscores for the primary and specific domains are estimated by solving the joint posterior distributions in Equation 3.2.2 and Equation 3.2.3 above from examinee response pattern, U . Particularly, the Gibbs sampling algorithm is used to iteratively samples θ from the full conditional distributions $\mathrm{U} \mid \theta,\xi$ and $\mathrm{U} \mid \theta,\zeta$ for the BF-C and BF-PC models, respectively (Casella & George, 1992; Gelman et al., 2004; Brooks et al., 2011; Gamerman, 1997; de la Torre & Song, 2009). The fully Bayesian approach to the aforementioned conditional distributions can be written as

$$P(\theta \mid \mathrm{U},\xi) \propto |\mathrm{I}|^{1/2} exp\left(-\frac{1}{2}\theta'\mathrm{I}^{-1}\theta\right) P(\mathrm{U} \mid \theta) \tag{3.2.4}$$

or

$$P(\theta \mid \mathrm{U},\zeta) \propto |\mathrm{I}|^{1/2} exp\left(-\frac{1}{2}\theta'\mathrm{I}^{-1}\theta\right) P(\mathrm{U} \mid \theta). \tag{3.2.5}$$

When Gibbs sampling is applied to each of the model, sequences of random variables of subscores for each examinee, $\theta_i^1, \theta_i^2, \ldots, \theta_i^t$ are sampled for which drawingθ_i^t is depending on the previous draw, θ_i^{t-1}. The proposed models are multidimensional with primary and specific θs. For each of the iteration in the Gibbs sampler, each subset of subvector of $\theta_{i1}, \theta_{i2}, \ldots, \theta_{iD}$ is drawn simultaneously that is conditional on all other components of θ. The item characteristics function or be referred to as the item response surface can be defined as the following:

$$P\left(\theta_{id} \mid \theta^{t-1}_{i(-d)}, \mathrm{U}_i, \xi_j\right) \tag{3.2.6}$$

or

$$P\left(\theta_{id} \mid \theta^{t-1}_{i(-d)}, \mathrm{U}_i, \zeta_j\right). \tag{3.2.7}$$

The above equations are described as at each state of the Gibbs sampling, dth specific content domain subscore of ith examinee, θ_{id}, is estimated conditionally on all other specific content domain subscores including the primary subscore as well as on item parameters from each of the bi-factor MIRT model. Thus, inferences regarding subscores for primary and specific content domains can be made from the posterior distribution.

Subscores are then be estimated from the multidimensional Bayesian expected a posteriori (EAP-M) by taking the means of the posterior distributions. In general, subscores for bi-factor compensatory and partially compensatory MIRT models can be computed from its EAP-M, respectively, as

$$\tilde{\theta}_{id} = E\left[\theta_{id} \mid \mathrm{U}_i, \xi_j\right] \approx \frac{1}{T-b}\sum_{t=b+1}^{T}\theta^t_d \tag{3.2.8}$$

or

$$\tilde{\theta}_{id} = E\left[\theta_{id} \mid \mathrm{U}_i, \zeta_j\right] \approx \frac{1}{T-b}\sum_{t=b+1}^{T}\theta^t_d, \tag{3.2.9}$$

where b is the total discarded samples from the Gibbs sampling of MCMC. $t = b+1, b+2, \ldots, T$ and T is the total number of iterations. Therefore, the posterior for the subscores estimates are computed from $T-b$ MCMC samples.

3.3 Simulation Study Design

Two simulation studies are designed, each for the bi-factor compensatory and bi-factor partially compensatory MIRT models respectively. Both models have the same simulation conditions. Many present studies on the bi-factor model and the multidimensional IRT focused on the accuracy of parameters estimation from these models. Several manipulated factors are observed to determine the generalizability of the obtained results are, amongst which, sample size, test length, number of items in each subscale, number of subscales and item discrimination patterns (DeMars, 2005; Thissen & Edwards, 2005; Edwards & Vevea, 2006; Edwards, 2010; Haberman, 2008; Shin, 2007; Tao, 2009; Yen, 1987; Deng et al., 2008; Gibbons & Hedeker, 1992; Gibbons et al., 2009; Immekus & Imbrie, 2008; Li & Rupp, 2011; Reise et al., 2010; Thompson, 2006).

3.3.1 Independent Variables

The primary focuses of this study are to examine two models of the bi-factor 3PL MIRT - compensatory and partially compensatory. The independent variables are two different test lengths and five levels of discrimination parameter. These two factors are to be considered as important in test that is developed from a multidimensional structure and clearly assessed abilities in primary and specific content domains.

3.3.1.1 Type of Item and Test Lengths

Only dichotomous or binary scored item, that is 0 for incorrect and 1 for correct, was be considered. Multiple-choice question, “right” or “wrong” , and “yes” or “no” items are all considered as dichotomous or binary items. This true data matrix of the item responses, that represents as U, was generated from the Monte Carlo simulation from which the true models are either the bi-factor compensatory 3PL MIRT model or the bi-factor partially compensatory 3PL MIRT model. In other words, the true item parameters, person or examinee parameter and item responses are generated from Equation 3.1.1 - Equation 3.1.2 or Equation 3.1.3 - Equation 3.1.4.

Three dimensions of specific content domains were examined by crossing two different test lengths - 30 and 60 items. For the 30-item test, each specific content domain has 10 items, and the 60-items test has 20 items in each specific content domain. The bi-factor 3PL MIRT models in this study have a total of four dimensions with two different test lengths (i.e. 30 or 60 items). The fixed factor for both simulation studies are the number of examinees, which is fixed to 1,500 people. This sample size is considerably a suffice number of examinees for parameters estimation of the studied models.

3.3.1.2 Item and Examinee Characteristics

For the primary and specific dimension of content domains, the person or examinee θ vector is randomly drawn from a multivariate normal distribution with a zero mean vector and an identity variance-covariance matrix, that is $\theta \sim MVN(0,I)$. θs range from -3 to 3 to posit from low ability to high ability level for both primary and specific content domains. The vector of item discrimination parameters, a, indicates rate of the probability of correct response changes from dimension to dimension in the θ continuum. The a-parameters are bounded by zero and typically distributed as log-normal with zero means and different levels variances or as uniform with a minimum of .2 and a maximum 2.0.

Different levels of the discrimination parameters, a_{jd}, reflect low to high discrimination between primary and specific abilities within a test. For this study, the discrimination parameter for the primary dimension distributed as uniform with a minimum of .75 and a maximum of 1.25. Five discrimination levels for the specific content domains dimensions were examined. Essentially, item discrimination of the specific content domains are equal, higher or lower than the primary dimension. Specifically, two levels of item discrimination parameters in the specific content domains are considered lower than the item discrimination in the primary dimension. The lowest level of item discrimination has a minimum of .25 and a maximum of .75. The next level of item discrimination has a minimum of .50 and a maximum of 1.00. These levels of item discrimination show that subscores in the specific content domains are less discriminative or informative than the primary

Table 3.1:
Simulation Condition for the Bi-factor Compensatory and Partially Compensatory 3PL MIRT Models

Model	*N*	*d*	Test Lengths	Discrimination Levels
Bi-factor Compensatory	1,500	1, 2, 3, 4	30 (10-10-10)	1.50, 1.25, 1.00, .75, .50
MIRT Model (BF-C)	1,500	1, 2, 3, 4	60 (20-20-20)	1.50, 1.25, 1.00, .75, .50
Bi-factor Partially Compensatory	1,500	1, 2, 3, 4	30 (10-10-10)	1.50, 1.25, 1.00, .75, .50
MIRT Model (BF-PC)	1,500	1, 2, 3, 4	60 (20-20-20)	1.50, 1.25, 1.00, .75, .50

Note. BF-C: Bi-factor Compensatory MIRT model, BF-PC: Bi-factor Partially Compensatory Model, *N*: Number of examinees, *d*: Dimension of primary and specific domains.

subscore. Also, when item discrimination parameter in the specific dimensions has equal size to the discrimination with the primary dimension, that is between .75 and 1.25, that means both specific content domains and primary domain are equally discriminating. Finally, item discrimination for the specific content domains are considered higher than the primary dimension. The size of this larger item discrimination is between 1.00 and 1.50 or between 1.25 and 1.75. These levels of item discrimination explain that subscores from the specific content domains are importance of informative value that the subscores represent distinct examinee's performance in each specific area of content domain besides the primary ability. Consequently, to consider the bi-factor model specification, there are five confirmatory structure matrices, S_1, S_2, S_3, S_4 and $, S_5$ which defined to reflect five levels of item discrimination parameters.

For the compensatory model item difficulty, b_j, for jth item was randomly drawn from $N(0,1)$. For the partially compensatory model, the item difficulty, b_{jd}, for jth item in dimension d was randomly drawn from $N(0,1)$. Difficulty parameters for both models are not primary focus in this simulation study, and thus the variances of b-parameters in both models are fixed to 1 and it can be left as parameters to be estimated without constraints. Typically, b-parameter is ranged from -2 to 2 to reflect very easy items to very difficult items. The pseudo-guessing, c_j, is bounded by 0 and 1, and typically be randomly drawn from a uniform distribution with a minimum of 0 and a maximum of .30.

Table 3.1 summarizes simulation conditions for this study. There were 2(test lengths) × 5(levels of discrimination) × 2(bi-factor MIRT models) = 20 simulation conditions considered in this study.

Examinee and item parameters are all be randomly drawn in independent trials. Because randomness involved in the generation of data via Monte Carlo simulation, true data for each condition is sampled for 50 times. To generate all the data, an R GUI programming language is developed for this study for each of the proposed model. The ability parameters for all dimensions, $\theta_{i1}, \theta_{i2}, \theta_{i3}, \theta_{i4}$, are randomly drawn from a multivariate standard normal distribution (*MVN*), and each examinee has a minimum of two orthogonal latent abilities for primary and specific content domains. In summary, true item and examinee parameters are generated from the following distributions:

Theta: $\theta \sim MVN(0, \mathrm{I})$ where $0 = \{0, 0, 0, 0\}'$ and $\mathrm{I} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}_{4\times4}$

Discrimination Parameter:

- Primary dimension: $a_{j1} \sim Uniform(.75, 1.25)$
- Specific Content Domain: $a_{j(d-1)} \sim Uniform(.25, .75)$, $a_{j(d-1)} \sim Uniform(.50, 1.00)$, $a_{j(d-1)} \sim Uniform(.75, 1.25)$, $a_{j(d-1)} \sim Uniform(1.00, 1.25)$, and $a_{j(d-1)} \sim Uniform(1.25, 1.75)$.

Difficulty Parameter:

- Compensatory Model: $b_j \sim Normal(0, 1.00)$
- Partially Compensatory: $b_{jd} \sim Normal(0, 1.00)$

Pseudo-guessing Parameter: $c_j \sim Uniform(.10, .30)$

3.3.1.3 Item Responses

The item responses for both primary and specific content domains are generated from the probability of correct response function, p, from Equation 3.1.1 or Equation 3.1.3. Based on Bernoulli trials, there are two outcomes, correct $(u = 1)$ and incorrect $(u = 0)$ and $0 < p < 1$. For any random variable, say z, that is drawn from a uniform distribution with a minimum of zero and a maximum of 1.0. If $z < p$ then the item response is equal to 1 for correct response, and if $z \geq p$ then the item response is equal to 0 for incorrect response. Therefore, $P(u = 1) = p$ and $1 - p$, respectively, represents probability of correct and incorrect response. There were1500 × 30 or 1500 × 60 data matrix generated from both BF-C and BF-PC models.

3.4 Estimation Procedures

There are 10 simulation conditions for each model and of which contains true values of the item and examinee parameters and a matrix of item responses. There are four dimensions of abilities with three specific dimensions of content domains, where there are 10 or 20 items in each of the specific content domain. Subscores for primary and specific content domains are estimated from the bi-factor compensatory (Study 1) or partially compensatory (Study 2) 3PL MIRT models using an R GUI package called R2OpenBUGS (Sturtz et al., 2005). The R2OpenBUGS package allows for running OpenBUGS (latest version: 3.2.1), and is employed to performing parameter estimation using the Bayesian inference with the Markov Chain Monte Carlo (MCMC) method using the appropriate Gibbs sampling algorithm. The package runs OpenBUGS interactively from within R.

3.4.1 Prior Distributions

Prior knowledge for multidimensional IRT parameters for both compensatory and partially compensatory for MCMC simulation is less clear when latent ability is interact with item characteristics in more than one dimension. As a rule of thumb, normal priors can be used for greater generalizability. Beguin & Glas (2001) recommended a normal distribution centered at means of 0 and a more diffuse variances of 2 for the (positive) discrimination parameter. Whereas, TESTFACT used even less informative normal priors with means of 0 and variances of 4 as the default. For thetas, multivariate normal priors with means of 0 and an identity variance covariance matrix can be used, which is also was chosen to align with the definition of the bi-factor specification for both compensatory and partially compensatory models that abilities are orthogonal. For generalizability less informative priors, normal priors for b-parameter with means of 0 and variances of 2 were suggested in Beguin & Glas (2001) and Edwards (2010). Users can choose different set of priors for pseudo-guessing parameters. One suggestion from Edwards (2010) is to set the prior for c-parameter in the 3PL models to a Beta distribution, and the number of response alternatives from a dichotomous item can be used to locate the c-parameter in MCMC simulation as well as to influence the c-parameter to converge to its reasonable target distribution.

In this study, to incorporate Bayesian approach with the Gibbs sampling for the estimation of subscores from Equation 3.2.2 and Equation 3.2.3, prior distributions are defined for item and person parameters with noninformative priors as follow:

Ability parameter, theta: $\theta \sim MVN(0,\mathrm{I})$ where $0 = \{0,0,0,0\}'$ and $\mathrm{I} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}_{4\times4}$.

Discrimination parameter: $a_{jd} \sim Normal\,(0, pr.a)\; T\,(0,\,)$ where $T\,(0,\,)$ means truncation for positive values, $\left(a_{jd} > 0\right)$.

Difficulty parameter: $b_j \sim Normal\,(0, pr.b)$ or $b_{jd} \sim Normal\,(0, pr.b)$.

Pseudo-guessing parameter: $c_j \sim Beta\,(20, 80)$.

Here, *pr.a* and *pr.b* are the precision of the hyperparameters for the variances of a- and b-parameter, respectively, which the variances were set to be equal as $\sigma_a^2 = 4.0$ and $\sigma_b^2 = 4.0$. This was set to follow the default in TESTFACT. Thus, the precision of the MCMC simulation for these parameters will be $pr.a = 1/\sigma_a^2$ or $pr.b = 1/\sigma_b^2$, respectively. Considering that a dichotomous item in a test has five answer options, thus the probability of guessing an option correctly is equal to .20 (i.e one of five). Therefore, to have a mean of .20 in c-parameter, the prior was controlled to follow a beta distribution with a minimum of 20 and a maximum of 80. Finally, priors for thetas were assigned to a multivariate normal with means of 0 and a covariance matrix equal to the identity matrix.

3.4.2 MCMC with Gibbs Sampling

Below is an outline of the MCMC algorithm with Gibbs sampling for subscores estimation, by focusing on the primary and specific dimensions of the ability parameters:

1. Select the initial values for each examinee's subscores at random and following a multivariate normal distribution as $\theta_i^t \sim MVN(0, \mathrm{I})$.

2. At iteration t, draws the candidate value $\theta_i^\star$ from previous state as $\theta_i^\star \sim MVN\left(\theta_i^{t-1}, \mathrm{I}\right)$ and define the next values for the subscores as $\theta_i^{t+1} = \left\{\theta_{i1}^{t+1}, \theta_{i2}^{t+1}, \theta_{i3}^{t+1}, \theta_{i4}^{t+1}\right\}$ where each subscore drawn from the following steps:

   (a) samples θ_{i1}^{t+1} from $p\left(\theta_{i1} \mid \theta_{i2}^t, \theta_{i3}^t, \theta_{i4}^t, \mathrm{U}, \xi_j\right)$ or $p\left(\theta_{i1} \mid \theta_{i2}^t, \theta_{i3}^t, \theta_{i4}^t, \mathrm{U}, \zeta_j\right)$

   (b) samples θ_{i2}^{t+1} from $p\left(\theta_{i2} \mid \theta_{i1}^t, \theta_{i3}^t, \theta_{i4}^t, \mathrm{U}, \xi\right)$ or $p\left(\theta_{i2} \mid \theta_{i1}^t, \theta_{i3}^t, \theta_{i4}^t, \mathrm{U}, \zeta_j\right)$

   (c) samples θ_{i3}^{t+1} from $p\left(\theta_{i3} \mid \theta_{i1}^t, \theta_{i2}^t, \theta_{i4}^t, \mathrm{U}, \xi_j\right)$ or $p\left(\theta_{i3} \mid \theta_{i1}^t, \theta_{i2}^t, \theta_{i4}^t, \mathrm{U}, \zeta_j\right)$

   (d) samples θ_{i4}^{t+1} from $p\left(\theta_{i4} \mid \theta_{i1}^t, \theta_{i2}^t, \theta_{i3}^t, \mathrm{U}, \xi_j\right)$ or $p\left(\theta_{i4} \mid \theta_{i1}^t, \theta_{i2}^t, \theta_{i3}^t, \mathrm{U}, \zeta\right)$

3. Return to Step 2.

In MCMC simulation, the subscores candidate, $\theta_i^\star$, is accepted with the following probability

$$p\left(\theta_i^{t-1}, \theta_i^\star\right) = min\left\{1, \frac{P\left(\mathrm{U}_i \mid \theta_i^\star\right) P\left(\theta_i^\star \mid \theta_i^t\right)}{P\left(\mathrm{U}_i \mid \theta_i^{t-1}\right) P\left(\theta_i^{t-1} \mid \theta_i^t\right)}\right\} \tag{3.4.1}$$

3.4.3 Checking Model Convergence

The number of MCMC samples and its discarded samples are important for model convergence as to ensure the overall MCMC estimation for each parameter in the studied model is converged to a stochastic solution. OpenBUGS outputs several plots for convergence diagnosis, among which, time series or trace plots, and kernel density plot. In this study, the convergence of the MCMC estimation was monitored for each parameter, and diagnostic plots provided by R2OpenBUGS package were examined and recorded.

Patz & Junker (1999) suggested that several chains would lead to a small Markov Chain (MC) standard error. This study decided to have two simulated chains for each monitored parameter from both bi-factor MIRT models. For more inspections of the model convergence, Gelman et al. (2004) and Gelman & Shirley (2011) proposed *R*-statistics by comparing within and between variances from the simulated chains. *R*-statistics is provided by R2OpenBUGS and it is called "*Rhat*", that is the estimated values of the scale reduction or shrink factors, and sometimes referred to as the Markov Chain variance that is computed by averaging the variances of the sub-chains over the number of sub-chains (Carlin & Louis, 2000). Thus, *Rhat* is used to assess the magnitude of cross-chains variance to total variance that can be mathematically expressed as

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{var}^{+}(\psi \mid U)}{W}},$$

where $\hat{var}^{+}(\psi \mid U) = \frac{T-1}{T}W + \frac{1}{T}B$, T is the number of MCMC samples, W is the average within-chain variance, B is the between-chain variance, ψ is the vector of all monitored parameters, and U is the data matrix. Note that,$\sqrt{\hat{R}} \to 1.0$ as $T \to \infty$ and approximate convergence has reached when $\sqrt{\hat{R}} < 1.20$ for all parameters (Gelman, 1996; de la Torre & Hong, 2010). If *Rhat* is greater that 1.20, further simulations should improve convergence. Results of *Rhat* for all simulation conditions are illustrated in Chapter 4.

3.5 Evaluation Criteria

This section includes two subsections for the evaluation of the simulation study and the subscores estimated from the studied models. In the first subsection, evaluation of the simulation study is divided into two parts where the first part covers the parameter recoveries of the simulation condition, and the second part demonstrates Bayesian approach for models comparisons and selections. Then, methods for evaluating reliability and classification of the estimated subscores are discussed.

3.5.1 Simulation Evaluation

In this section, methods to evaluate parameter recoveries from the studied models are discussed followed by the Bayesian approach for MCMC simulation evaluation.

3.5.1.1 Parameter Recoveries

Bias, absolute bias, root mean square error (RMSE) and standard error of estimate (SEE) are computed for each parameter in the models. Bias and RMSE are further examined to assess the accuracy of parameter estimated over the 50 replications of each of the simulation condition. Pearson's correlations are computed and plotted to summarize the relationships between the estimated subscores and the true abilities for primary and specific dimensions.

Bias, RMSE and SEE for each item or person parameter,(δ), are computed for all 50 replications by the following formulae

$$Bias\,(\delta) = \frac{1}{n}\sum_{n=1}^{N}(\delta_e - \delta_t)$$

$$Absolute Bias\,(\delta) = |Bias| = \frac{1}{n}\sum_{n=1}^{N}|\delta_e - \delta_t|$$

$$RMSE\,(\delta) = \frac{1}{n}\sqrt{\sum_{n=1}^{N}(\delta_e - \delta_t)^2}$$

$$SEE\,(\delta) = \frac{1}{n}\sqrt{\sum_{n=1}^{N}\left(\delta_e - \bar{\delta}_e\right)^2} \text{or} \, SEE = \sqrt{RMSE\,(\delta)^2 - Bias\,(\delta)^2},$$

where n is the number of items or examinees and there are 50 replications are considered for each simulation condition. δ_e is the estimated values and δ_t is the true generated values at the nth item of examinee and $\bar{\delta}_e$ is the expected values of the estimates, that is $\bar{\delta}_e = \frac{1}{n}\sum\hat{\delta}_e$.

The dependent variables of interest in this simulation study are bias and RMSE. For each of the studied models, compensatory or partially compensatory, repeated measures ANOVA, series of ANOVA and ANOVA were used to test any significant differences in the bias, and RMSE over the simulation conditions for all item and person parameters. That is to test the interaction effect between test length and discrimination level, main effect of test length and main effect of discrimination level. Results for all items parameters and thetas are presented in Chapter 4.

3.5.1.2 Bayesian Complexity and Fit

In addition to the examination of *Rhats* for model convergence diagnostics, Spiegelhalter et al. (1998), Spiegelhalter et al. (2002), Gelman et al. (2004) and Gelman & Shirley (2011) suggest Bayesian deviance and *deviance information criterion* (DIC) as the measure of fit and complexity or adequacy from model that best explain the observed data. DIC can be computed from the "complexity" of the model, that is by the *effective number of parameters*, and the "fit" of the model , that is called deviance. This procedure is analogous to the computation of residual in classical model-fit assessment. DIC may be used for model comparisons and selection, and it can be computed as

$$DIC = \bar{D} + p_D,$$

where

$\bar{D}$ ("Dbar") is the *deviance* and is posterior expectation of the deviance summarizes the fit of the studied model and $\bar{D} = E_{\theta|\mathrm{U}}[D] = -2logL(\mathrm{U} \mid \theta)$; p_D is the effective number of parameters and is computed as $\bar{D} - D\left(\bar{\theta}\right)$. $D\left(\bar{\theta}\right)$ ("Dhat") is the deviance at the posterior expectations which is the standardization of the -2log(likelihood). DIC can be shown as the approximation of the *Akaike*'*s information criterion* (AIC) as

$$AIC = \bar{D} + 2p_D.$$

Kelly & Curtis (2011) and Spiegelhalter et al. (2002) recommend another Bayesian fit index for models justification that is referred to as the *Bayesian information criterion* (BIC). BIC can be computed as

$$BIC = \bar{D} + p_D log(N),$$

where N is the number of observations. Smaller value of *deviance*, DIC, AIC and BIC indicate a better-fitting model.

3.5.2 Subscores Evaluation

This study recommended that EAP subscores estimated from bi-factor compensatory and partially compensatory models be evaluated from the perspective of reliability and validity. Reliability and validity of the estimated EAP subscores obtained from each condition were compared to the true generated subscores across simulation conditions. For reliability, Pearson's correlation was used to evaluate the accuracy of the estimated primary and specific content domains subscores. Also, Bayesian marginal reliability for each EAP subscore are computed and compared. Finally, classification of the primary and specific content domain subscores are evaluated from the subscore separation index defined in this study. This measure might also be elaborated as validity of the subscore estimates.

3.5.2.1 Subscore Correlation

In general, the relationships between true and estimated subscores for each simulation condition was obtained from Pearson's correlation coefficient by the following:

$$r = \frac{1}{nReps} \sum_{r=1}^{nReps} cor(\theta_e, \theta_t)^2,$$

where r is the vector of average correlation. θ_e and θ_t, respectively, represent the vector of estimated and true subscores for primary and specific content domains, and *cor* is the correlation function. Higher correlation or reliability of each subscore, r_d, shows that the subscore estimates recover the true subscore well.

For the Bayesian subscore estimates, reliability was evaluated by a marginal reliability estimate based on the ratio of estimated true subscore variance to the estimated subscore variance. Conventionally, this reliability is defined by coefficient alpha (Wainer et al., 2001; Skorupski & Carvajal, 2009). For the Bayesian perspective of the multidimensional models, the marginal reliability for subscores estimation was computed from the multidimensional Bayesian expected a posteriori (EAP-M) as the ratio of the variance of the ability estimates to the sum of this variance plus the average squared standard error. This study applied a method proposed by de la Torre & Patz (2005) that is known as the Bayesian marginal reliability for each EAP subscore estimate, and is analogous to the coefficient alpha, can be obtained from

$$\hat{\rho}_d = \frac{var\left(\hat{\theta}_d\right)}{var\left(\hat{\theta}_d\right) + \overline{Pvar\left(\hat{\theta}_d\right)}} \equiv \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}, \tag{3.5.1}$$

where $var\left(\hat{\theta}_d\right)$ is the variance of the dth EAP subscore for N examinees, that is $var\left(\hat{\theta}_d\right) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\theta}_d - \overline{\theta}\right)^2$, and $\overline{Pvar\left(\hat{\theta}_d\right)}$ is the marginal posterior EAP subscores variance of the primary or specific ability estimates, that is $\overline{Pvar\left(\hat{\theta}_d\right)} = \frac{1}{N}\sum_{i=1}^{N} SE_{\hat{\theta}_i}^2$. Higher Bayesian marginal reliability demonstrates higher subscores reliability from the studied models.

3.5.2.2 Subscore Separation Index

Subscore separation index (SSI) was used to quantify the different between subscores within a person. SSI was proposed by Tate (2004) and empirically examined in Tao (2009). Subscore separation index can be computed from a ratio of the differences between a pair of the estimated subscores to the sum of the standard errors of the two subscores. The percentage of this ratio over examinees that is greater than 1.0 for all pairs of subscores is the measure of SSI. This study defines a mathematical expression of the SSI as follows

$$SSI_i = \frac{|\hat{\theta}_{(d-1)} - \hat{\theta}_1|}{SE_{\hat{\theta}_{(d-1)}} + SE_{\hat{\theta}_1}}, \tag{3.5.2}$$

where $(d-1)$ represents any subscores other than the first subscore, which is the primary subscore. Thus, for each examinee *i*, *SSI* is defined as the ratio between an absolute different between a pair of subscores - subscore in one specific content domain and subscore in the primary ability, to the total standard errors from both domains. Therefore, separation or classification of subscores from the specific content abilities from the primary ability can also be computed as an index of *SSI* by the following

$$SSI = \frac{1}{N}\sum_{i=1}^{N}\left(SSI_i\right).$$

For this simulation study, *SSI* is computed as the proportion of subscores separation between primary and specific content domains and it is determined from the perspective of Bayesian framework of the EAP subscores from the specific content domains. Thus, using the same above expression, the computation for *SSI* was performed such that: $\left|\hat{\theta}_{(d-1)} - \hat{\theta}_1\right|$ represents the absolute different between a pair of EAP subscores. $SE_{\hat{\theta}_{(d-1)}}$ and $SE_{\hat{\theta}_1}$, respectively, are the posterior standard error for the specific and primary dimension. The posterior standard error can be obtained from the standard deviation of the posterior distribution. To investigate the important inference about the EAP subscores estimates of the specific content domains over the primary score estimate, the EAP subscores estimates are always paired with the primary score estimate. *SSI* is averaged over examinees for each pair of subscores, and is primarily assessed at all levels of discrimination. It is expected that *SSI* would increase with the increasing discrimination level of the specific content domains. All *SSI*s are reported in Chapter 4 for both the bi-factor compensatory and partially compensatory MIRT models.

At each time *SSI* for each examinee is greater than 1.0, it will be counted to infer that subscores are important to be reported because the subscores are 1 level distinct from each other. The percentage of the counts over examinees for each simulation condition is referred to as *hit rate*. This study defines subscores hit rate as

$$HR\left(SSI\right) = P\left(SSI_i > 1.0\right) \times 100\% = \frac{1}{N}\sum_{i=1}^{N}\left(SSI_i > 1.0\right) \times 100\%.$$

The hit rate, $HR\left(SSI\right)$, may be used as an indicator to quantify about sensitivity of the subscores to be reported and to have practical important over the primary score. Thus, improvement, remedial or diagnostic actions about the examinee abilities on the specific content domains can be observed. The *SSI* and hit rate can also be used at which simulation condition(s) the estimated subscores produce important inferences over primary dimension. Thus, these indices are useful for setting a benchmark or cutting point to the development of tests with bi-factor multidimensional structures.

Chapter 4

Results

In this chapter, results from two simulation studies are presented. Results for Study 1 considered data sets simulated from the bi-factor compensatory multidimensional IRT model (BF-C), and results for Study 2 considered data sets simulated from the bi-factor partially compensatory IRT model (BF-PC). Each simulation study was designed to discuss research questions and hypotheses addressed in Chapter 1. The two factors in estimating subscores, reliability and classification that considered in the simulation studies were test length and discrimination level.

For both Study 1 (compensatory model) and Study 2 (partially compensatory model), the results presented in the subsection 4.2.2 and subsection 4.3.2 answer the first research question addressed in Chapter 1 of this dissertation. Furthermore, subsection 4.2.1, subsection 4.2.3,

subsection 4.3.1, and subsection 4.3.3 focus on the MCMC convergence diagnostics and Bayesian complexity and fit that answer the second research question. Finally, the last section in each study, that is subsection 4.2.4 and subsection 4.3.4, is devoted to answer the third research question. Discussion and conclusion from the results associated with this dissertation's research questions are presented in Chapter 5.

4.1 Parallel Computing Design for the Simulation Studies

For computing intensity of the Bayesian MCMC simulation, these simulations were conducted on the HPC cluster maintained by the Center for Research Methods and Data Analysis at the University of Kansas. This is a Rocks Cluster of Linux compute nodes running on Dell Power Edge 2950 Servers that have 16GB RAM and dual quad-core Intel Xeon processors. The simulation is managed by tools provided with R-2.14.2 and auxiliary packages, including parallel and R2OpenBUGS. Bayesian MCMC with Gibss sampling is conducted by OpenBUGS version 3.2.1.

There were 50 replications of analysis for each simulation condition in each model. Each run of the simulation is initiated by a set of random seeds. This design blends the advice of Chambers (2008) and L'Ecuyer et al. (2002). The L'Ecuyer random generator's stream can be subdivided into a large number of separate random number sequences, the starting point of each of which can be recorded and used to initiate a sequences of random numbers. That random generator passes stringent tests which assure that the random streams are uncorrelated and homogeneous. Before the simulations begin, an auxiliary program is run to collect the starting positions of each separate sub-sequence and those starting points and the sub-stream seeds are stored in a file for future use. In this dissertation, there were three random number streams for each run of the model. One stream is used to create the sample of examinees ability levels. In this design, the examinee abilities are supposed to be the same across all runs within each model, so the seed value is the same in each simulation of the model. This is to assure that each test is administered to the same set of students, and if we increase the number of students, then the first block of students will remain the same as more students are added. The other random streams are used to initialize the distributions for the test item parameters (i.e. difficulty, discrimination and pseudo guessing parameters) and item responses. When each run is initiated on the compute cluster, the program is designed to consult the saved set of random seeds, so that each run can be re-started and replicated exactly. More importantly, this design assured that there are not un-intended duplications of random number sequences that may make some runs of the model appear more similar than they ought to be at random.

For each study, an R program was written to generate true item and examinee parameters as well as item responses. OpenBUGS was used for MCMC simulation with Gibbs sampling to estimate the expected a posteriori (EAP) subscores for primary and specific ability dimensions. All item and examinee parameters as well as Bayesian convergence diagnostic criteria were monitored and estimated in OpenBUGS within the written R program.

The overall study examined a total of 20 simulation conditions. For each simulation condition, they were two parallel chains of the Markov Chain simulation that were observed. For the 30-item condition, the total number of MCMC iterations were 19,500 and the number of discarded iterations were 13,500 that guaranteed all of the parameters had converged. The chains were run using the same number of iterations and discarded iterations. Note that, there was a total of 6,180 number of monitored parameters for the BF-C model and 6,270 parameters for the BF-PC model in each of the MCMC simulation. This caused a very long simulation with a limitation computer memory and storage. Therefore, this study applied ×3 thinning to each chain and thus reduced the number of iterations to 6,500 and the number of discarded iterations to 4,500, which in turn only 2,000 MCMC samples for each chain were retained to make inference from the posterior distributions. The analysis for 50 replications in each studied simulation condition took about two and a half days and, thus, each run took an average of approximately less than one and one-quarter hours.

For the 60-item condition, to assure MCMC convergence, the total number of iterations were 24,000 and the number of discarded iterations were 18,000. The chains were also run using the same number of iterations and discarded iterations. For this test length, there was a total of 6,360 number of monitored parameters in the BF-C model and 6,600 parameters in the BF-PC model for each MCMC simulation, thus, ×3 thinning to each chain were done and that reduced the number of iterations to 8,000 and the number of discarded iterations to 6,000. Only 2,000 MCMC samples for each chain were also retained for this simulation condition. The 50 replications in each studied simulation condition for the MCMC analysis took a little bit longer than the 30-item condition. Four days of analysis were observed and, thus, each run took an average of approximately two hours to complete.

4.2 Results for Study 1

For this study, 10 sets of true data were generated from the bi-factor compensatory MIRT model (BF-C), specifically the M3PL model discussed from Equation 3.1.1 to Equation 3.1.2 in Chapter 3.

4.2.1 MCMC Simulation Diagnostics for Study 1

The Bayesian convergence criterion throughout this study was based on the multivariate potential scale reduction factor, that is called as Gelman-Rubin multivariate reduction factor or shrink factor (Gelman et al., 2004) and is denoted as $\hat{R}$. Plots of the average of $\hat{R}$ for item parameters on 30-item test and 60-item test at five discrimination levels are shown in Figure 4.1 and Figure 4.2, respectively. Whereas, plots of the average $\hat{R}$ for examinee (or person) parameters are illustrated in Figure 4.3 and Figure 4.4.

The x-axis has numbers representing simulation replications, and y-axis representing values of the $\hat{R}$ for item parameters or examinee parameters. For the item parameters, the solid black line is the $\hat{R}$ for a-parameter (discrimination), the dashed red line is for b-parameter (difficulty), and the dotted green line is for c-parameter (psuedo-guessing), and the solid black line in the examinee parameters represent the $\hat{R}$ for the latent abilities, thetas.

All of the plots of the Gelman-Rubin reduction factors for all parameters in the model showed all $\hat{R}s$ were less than 1.20. These results indicating that the output from all parallel chains of the MCMC simulations is indistinguishable that the cross-chains variance from the between-chain and within-chain variances estimates is unbiased. Therefore, all of the monitored item and examinee parameters have eventually converged to the stationary distribution which represents the joint posterior distribution as in Equation 3.2.4 in Chapter 3.

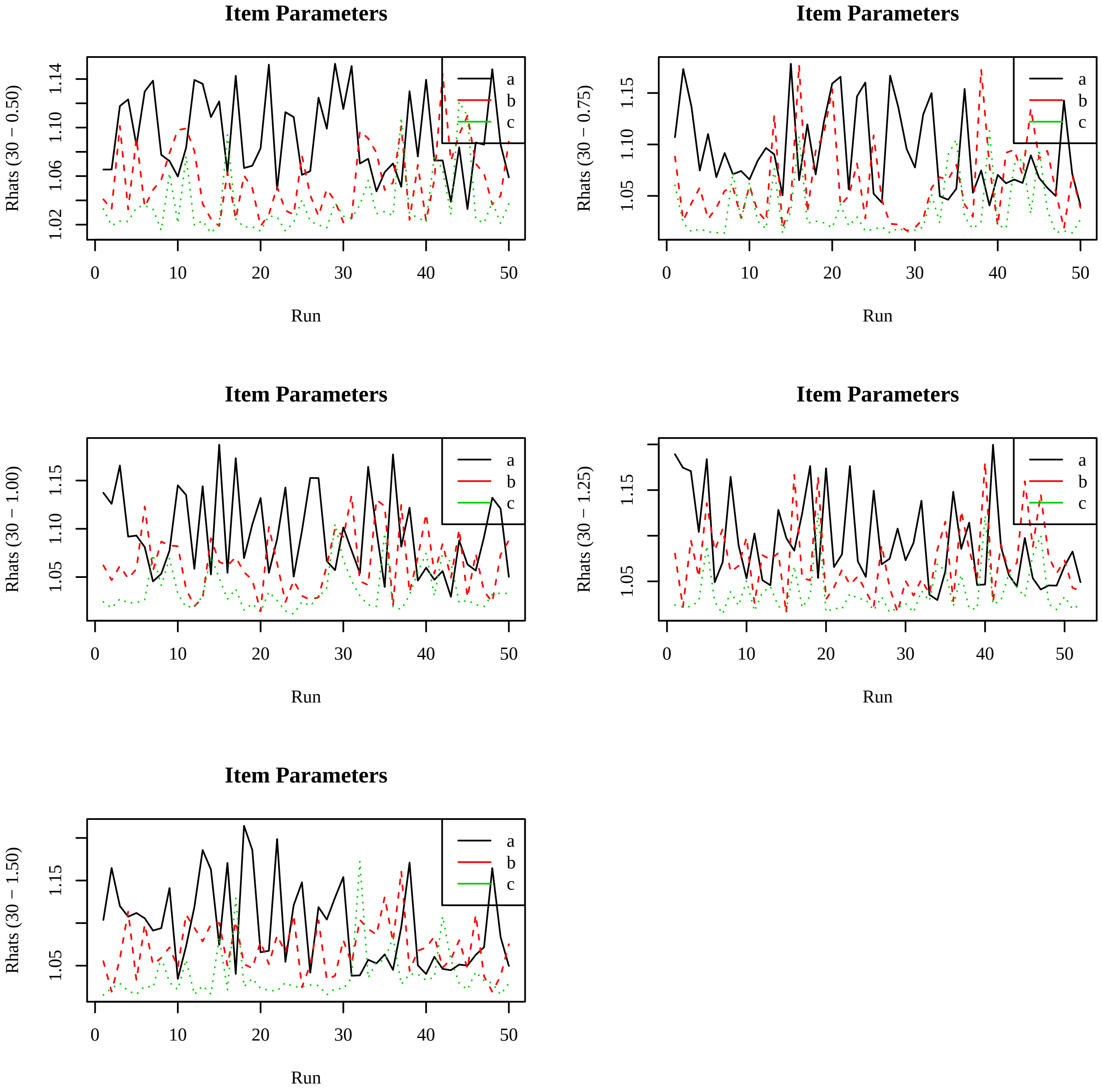


Figure 4.1: Rhats for Item Parameters in BF-C Model for 30-item Test

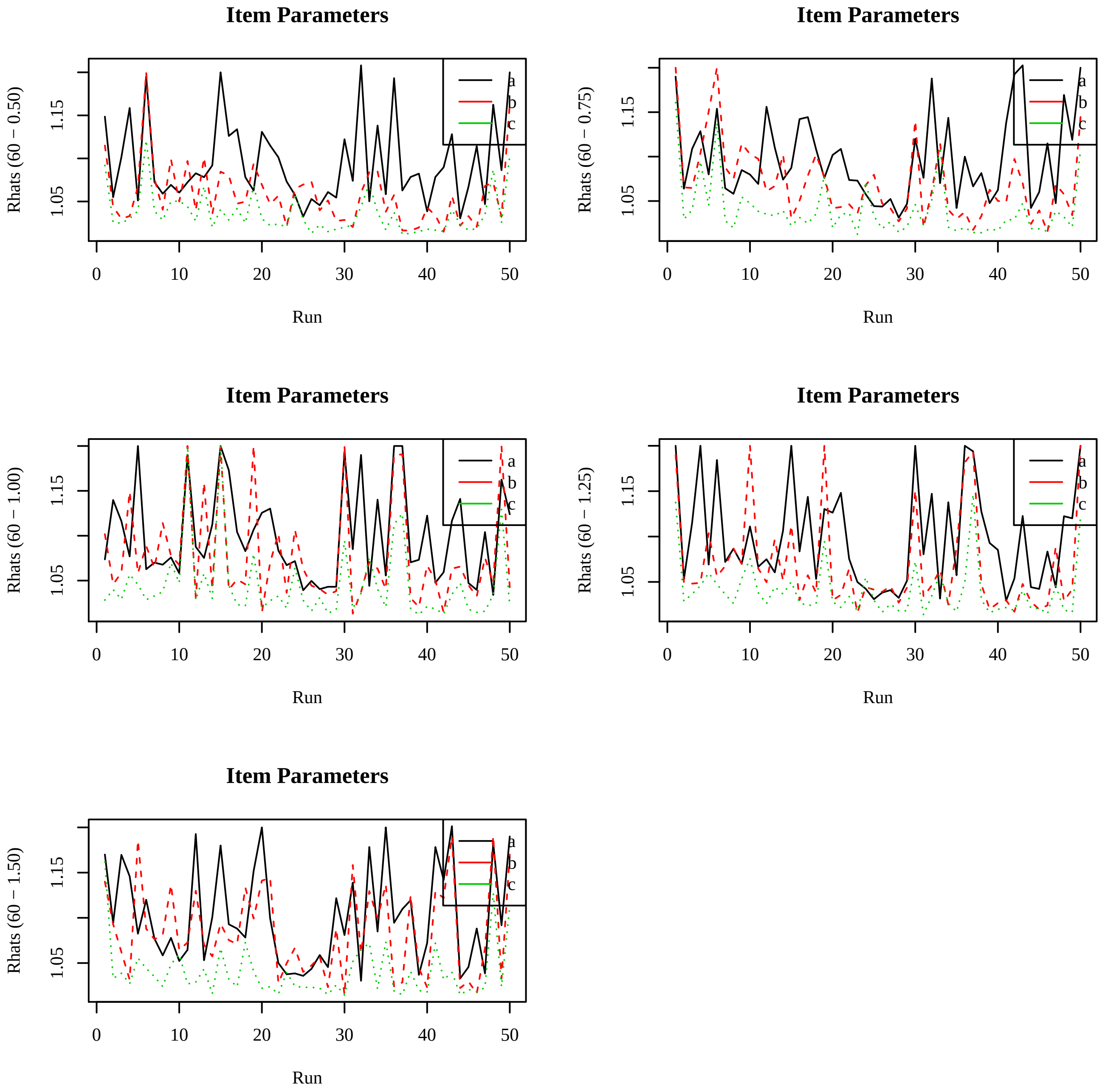


Figure 4.2: Rhats for Item Parameters in BF-C Model for 60-item Test

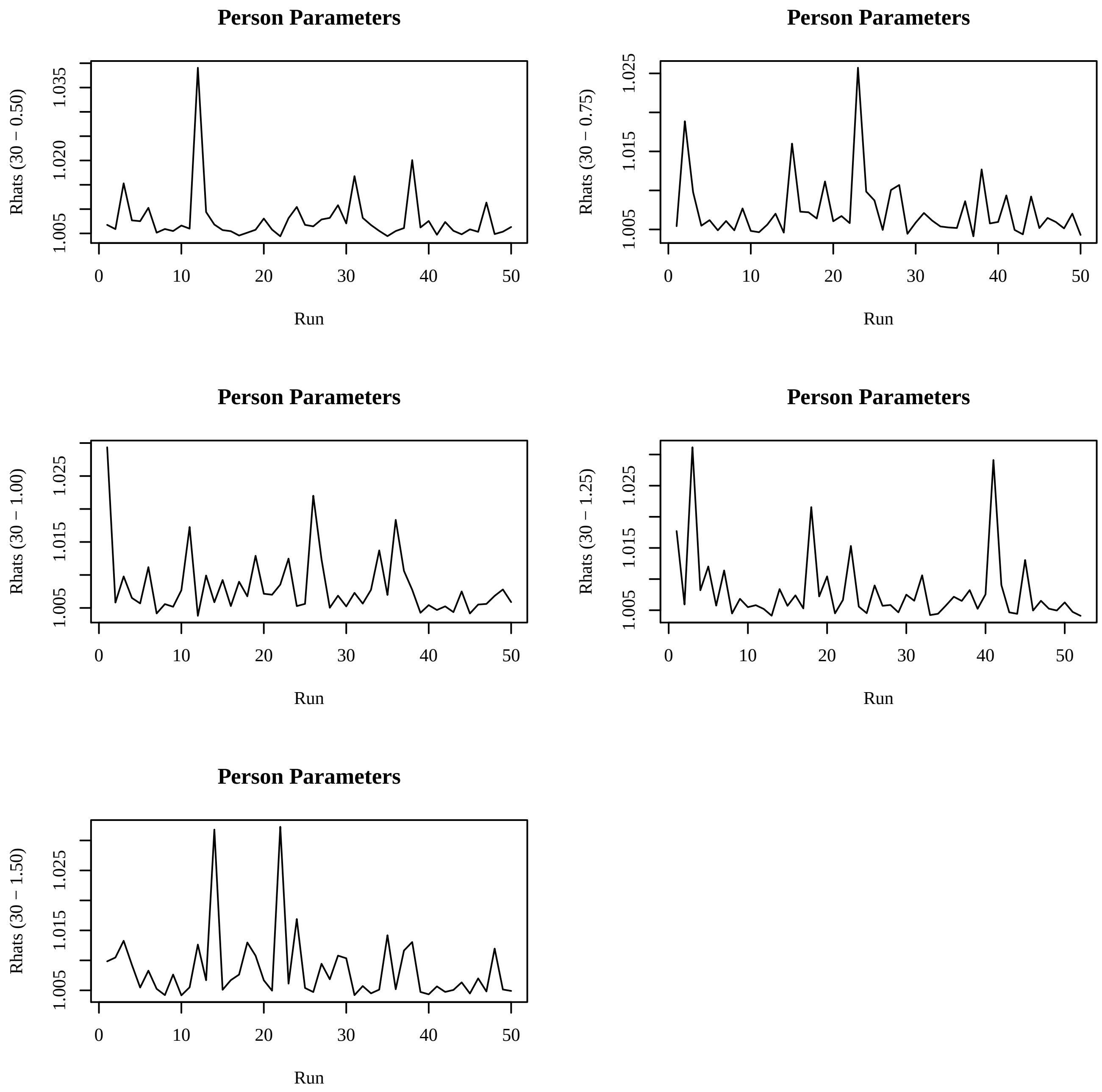


Figure 4.3: Rhats for Examinee (or Person) Parameters in BF-C Model for 30-item Test

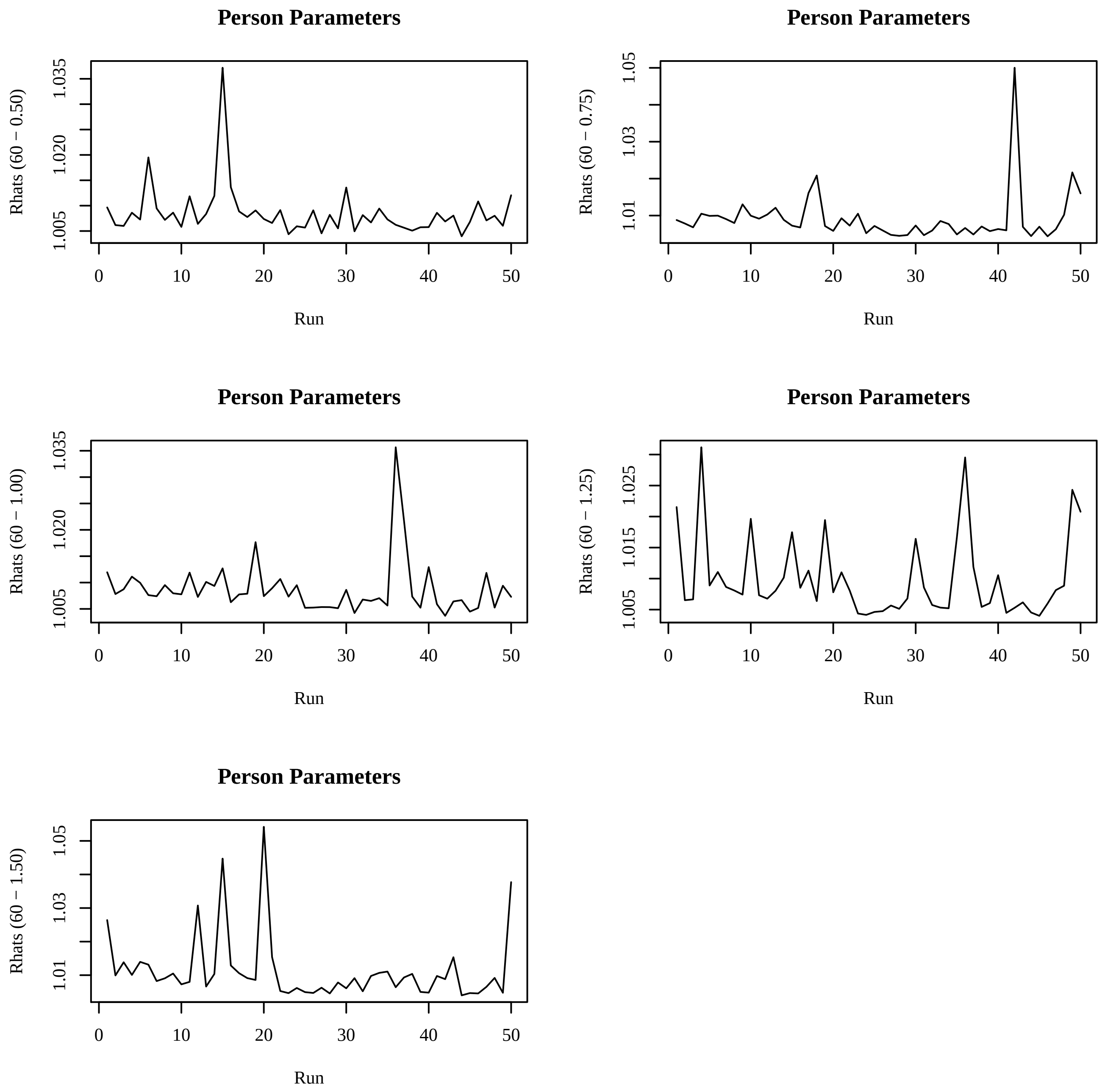


Figure 4.4: Rhats for Examinee (or Person) Parameters in BF-C Model for 60-item Test

Due to the large number of monitored parameters from all replications in each simulation condition, only one example of some of the trace and density plots are attached for this model (see Appendix A) . These plots are showed in Figure A.1 to Figure A.3 that are used for visual inspections for the model convergence diagnosis. The trace plot is a plot of the draw for iteration against the number of iterations. The trace plots for all selected parameters show good mixing as the draws from MCMC simulation from the two chains were moving around the same parameter space of the statistical distributions. Also, the curves from the Kernel density plots in each selected monitored parameter displayed smooth curves that their posteriors come from the desired posterior distributions.

4.2.2 Parameter Recovery for Study 1

This section describes solution to the first research question addressed in Chapter 1. Bias and RMSE for item and examinee parameters were assessed for simulation accuracy. Repeated measures analysis of variance (RM-ANOVA) using SAS PROC GLM procedure was used to evaluate bias and RMSE for item discriminations and examinee abilities (i.e. thetas) that have been estimated at four different ability dimensions. The between-runs factors for this procedure were two levels of test length (TL: 30 or 60 items) and five levels item discrimination (DL: .50, .75, 1.00, 1.25 and 1.50). For discrimination levels, there are two lower levels of discrimination (.50 and .75), one equally discriminate dimension (1.00), and two higher levels of discrimination (1.25 and 1.50) were studied. Whereas, the analysis of variance (ANOVA) was used to evaluate bias and RMSE for item difficulty and item pseudo-guessing.

4.2.2.1 Recovery of a-parameter

From Table A.1 in the Appendix A it follows that the bias of a-parameter in the primary dimension for the 30-item test were negative compared to the 60-item that were positively biased. Bias of a-parameter in the secondary ability dimensions were negative in the 30-item test but smaller than the 60-item test. For the first dimension, a-parameters were positively biased when the discrimination level of the specific ability dimensions was equal or lower than the primary dimension, and contrary, all a-parameter estimates were increasingly negatively biased when the level of discrimination levels of the specific ability dimension were higher than the primary dimension. However, in the secondary ability dimensions, all of the a-parameter estimates were all negatively biased, in which the a-parameters were less biased as the discrimination level decreased. These results are illustrated in 4.5a. 4.5b that also showed the results for the RMSE of a-parameter. Higher RMSE was observed for the conditions in shorter test and lower discrimination level. In other words, there were reduction in the error variance of the estimated a-parameters as the level of discrimination increases.

For the significant different tests of bias in a-parameters, RM-ANOVA was carried out. The Mauchly's sphericity test showed a significant result that bias and RMSE of a-parameter do not meet the sphericity assumption, $W < .001, \chi\,(5) = 5149.06, p < .001$ and $W = .058, \chi\,(5) = 1973.71$, $p < .001$, respectively. These results suggesting that the observed matrices do not have approximately equal variances and equal covariances. Thus, corrected RM-ANOVA F-tests are used such that Greenhouse-Geisser and Huynh-Feldt epsilon corrections were considered. The corrective coefficients were: Greenhouse-Geisser $\varepsilon_{bias} = .340$, $\varepsilon_{rmse} = .408$, and Huynh-Feldt $\varepsilon_{bias} = .340$, and $\varepsilon_{rmse} = .408$.

The multivariate approach for the within-runs and between-runs tests were then used to evaluate the bias and RMSE of a-parameter as a function of ability dimension. The null hypothesis is that bias or RMSE of a-parameter does not change across different ability dimensions. Table 4.1 and Table 4.2 summarize RM-MANOVA for bias and RMSE fo a-parameter. The Wilks' lambda (Λ) and Pillai's trace (V) are reported for each effect. The three-way interactions of ability dimension, test length and discrimination level for both bias and RMSE were not significant, $F_\Lambda\,(12, 1291.40) = .069, p = .763$ and $F_\Lambda\,(12, 1291.40) = .74, p = .716$. All two-way interactions and main effects were significant. These results suggest that test length and discrimination level, respectively, interacts with ability dimensions, $F_\Lambda\,(3, 488) = 38.56, p < .001$ and $F_\Lambda\,(12, 1291.40) = 2.16, p = .01$ for bias, and $F_\Lambda\,(3, 488) = 6.48, p < .001$ and $F_\Lambda\,(12, 1291.40) = 74.33, p < .001$

for RMSE. Therefore, bias and RMSE of a-parameter across different ability dimensions depends upon test length and discrimination level. In overall, bias and RMSE of a-parameter were substantially significantly changed with ability dimension, $F_{\Lambda}(3,488) = 203.42, p < .001$ and $F_{\Lambda}(3,488) = 8807.27, p < .001$.

For the between-runs effects, the interaction between test length and discrimination level was not significant for the bias of a-parameters, $F(4,490) = .160, p = .958, \eta^2 = .004$, and the interaction was significant for the RMSE of a-parameters with a very small effect size, $F(4,490) = 10.95, p < .001, \eta^2 = .05$. Both main effects of test length and discrimination level were significantly effects the RMSE of a-parameters, respectively $F(1,490) = 12.57, p < .001, \eta^2 = .056$ and $F(4,490) = 1825.11, p < .001, \eta^2 = .82$ for bias and respectively $F(1,490) = 59.08, p < .001, \eta^2 = .06$ and $F(4,490) = 438.08, p < .001, \eta^2 = .82$ for RMSE. Although there was a statistical significant different of the bias or RMSE of a-parameters between the 30-item and 60-item tests, the effect sizes were very small $\left(i.e.\, \eta^2_{bias} = .05, \eta^2_{rmse} = .06\right)$. Whereas, there were equally large effect sizes $\left(i.e.\, \eta^2_{bias} = .82, \eta^2_{rmse} = .82\right)$ on the significant different across five levels of discrimination on the bias and RMSE of a-parameters. Figure 4.5a and 4.5b depict these results on the bias and RMSE of a-parameters.

Table 4.1: *Multivariate Within-runs Effects for a-parameters in the BF-C Model*

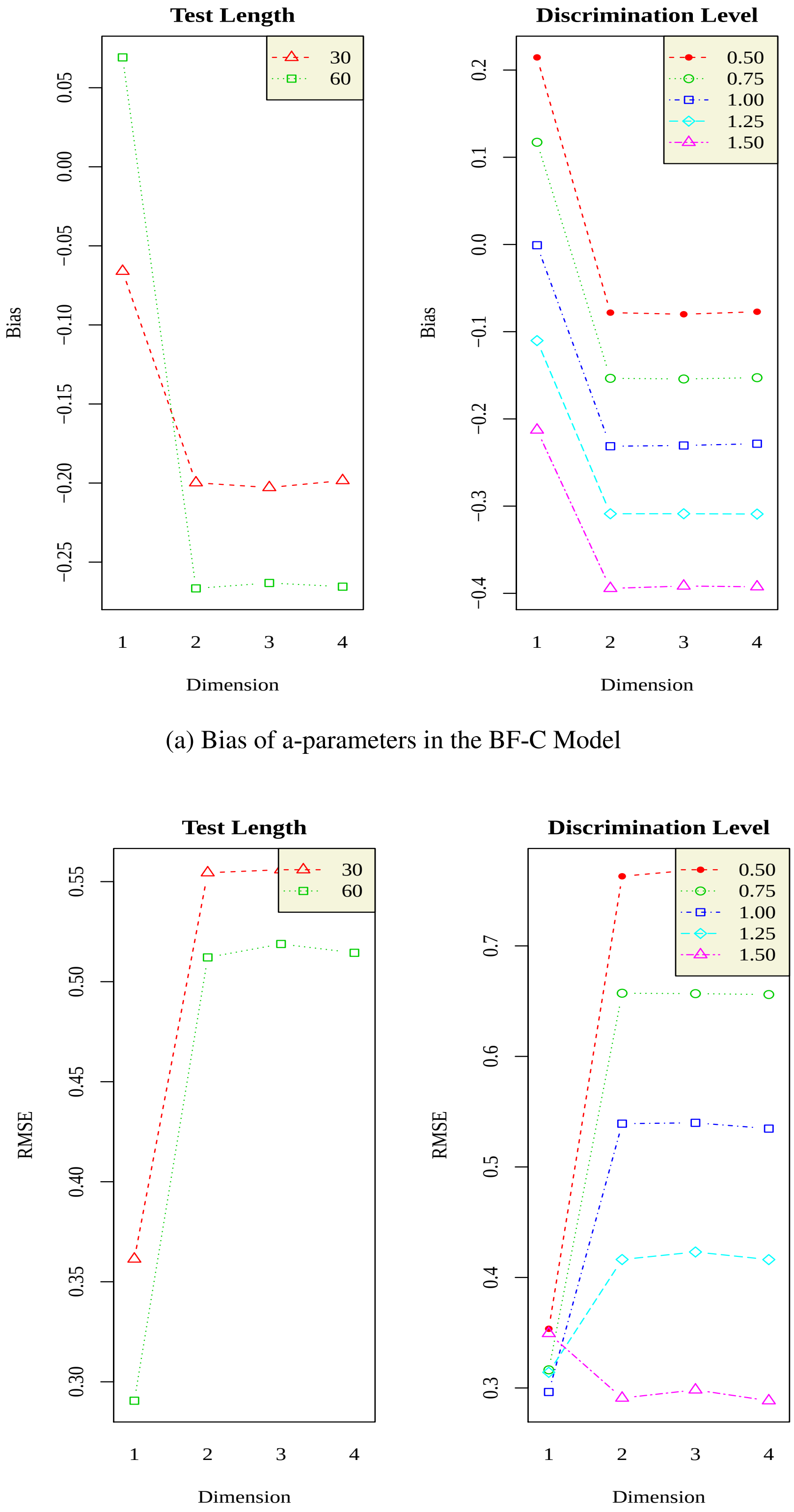
Effect	Λ	V	$F_\Lambda\,(df_1, df_2)$	$F_V\,(df_1, df_2)$
Bias				
Dimension*	.530	.470	203.42(3,488)	203.42(3,488)
Dimension×TL*	.856	.144	38.56(3,488)	38.56(3,488)
Dimension×DL**	.963	.037	$2.16(12, 1291.40)$	$2.14(12, 1470)$
Dimension×TL×DL	.988	.012	$.069(12, 1291.40)$	$.069(12, 1470)$
RMSE				
Dimension*	.241	.759	723.25(3,488)	723.25(3,488)
Dimension×TL*	.973	.027	6.48(3,488)	6.48(3,488)
Dimension×DL*	.348	.657	$74.33(12, 1291.40)$	$48.33(12, 1470)$
Dimension×TL×DL	.987	.013	$.74(12, 1291.40)$	$.74(12, 1470)$

Note: RMSE: Root mean square error, TL: Test Length, DL: Discrimination Level, df: degrees of freedom, *p<.001, **p<.01, Λ :Wilks' Lamdba, V : Pillai's trace.

Table 4.2: *Between-runs Effects for a-parameters in the BF-PC Model*

Effect	MS	F	df	η^2
Bias				
Test Length*	.153	12.57	1	.05
Discrimination Level*	10.32	1825.11	4	.82
Test Length × Discrimination Level	<.001	.16	4	.004
Error			490	
RMSE				
Test Length*	1.50	59.18	1	.06
Discrimination Level*	11.11	438.08	4	.82
Test Length × Discrimination Level**	.028	10.95	4	.05
Error			490	

Note: RMSE: Root mean square error, MS: Mean square error, df: degrees of freedom, *p<.001, **p<.01.



(a) Bias of a-parameters in the BF-C Model

(b) RMSE of a-parameters in the BF-C Model

Figure 4.5: Bias and RMSE for a-parameters in the BF-C Model

4.2.2.2 Recovery of b-parameter

For the difficulty parameter of this bi-factor confirmatory compensatory MIRT model, the effects of test length and discrimination level on bias and RMSE were examined using ANOVA from SAS PROC GLM procedure. Separate means and standard errors for b-parameter are presented in Table A.2 in Appendix A. The results from ANOVA in Table 4.3 show the interactions between test length and discrimination level for both bias and RMSE were not significant with very small effect sizes, $F\,(4,490) < .001, p = .997, \eta^2 < .001$ and $F\,(4,490) = .27, p = .602, \eta^2 = .001$, respectively. Also, with very small effect sizes, the effects of discrimination level on bias and RMSE were also not significant, $F\,(4,490) = .63, p = .428, \eta^2 < .001$ and $F\,(4,490) = 2.92, p = .08, \eta^2 = .004$. Only the effect of test length on bias and RMSE were significant, $F\,(1,490) = 14.07, p = .02, \eta^2 < .001$ and $F\,(1,490) = 32.39, p < .001, \eta^2 = .04$, respectively and the effects of test length on bias and RMSE of b-parameter were very small, $\eta^2_{bias} < .001$ and $\eta^2_{rmse} = .04$.

Figure 4.6 illustrates the interaction plots of the bias and RMSE of b-parameter. The b-parameter estimates were positively biased with RMSE ranged approximately between 1.5 and 2.7. The bias in the b-parameter of the 30-item test was increased when the level of discrimination increases from the lower to equal discrimination before the bias dropped down as discrimination levels increased.

There were unclear changes in the bias of b-parameter for the 60-item test. Moreover, the plots showed that the b-parameter in the shorter 30-items was more biased and RMSE was larger than the longer 60-items test. There were increasing patterns in the RMSE when the discrimination level increases both in the 30- and 60-item tests, and there was a slightly dropped in the RMSE at the highest level of discrimination in the 30-item test. These results indicating poor recoveries for b-parameter in the BF-C model.

Table 4.3: *Bias and RMSE of b-parameter in the BF-C Model*

Effect	*MS*	*F*	*df*	η^2
Bias				
Test Length*	7.598	14.07	1	.02
Discrimination Level	.340	0.63	4	<.001
Test Length × Discrimination Level	<.001	<.001	4	<.001
Error	.540		490	
RMSE				
Test Length*	76.565	32.39	1	.04
Discrimination Level	6.892	2.92	4	.004
Test Length × Discrimination Level	.643	.27	4	<.001
Error	2.364		490	

Note: RMSE: Root mean square error, MS: Mean square error, *df*: degrees of freedom, *p<.001.

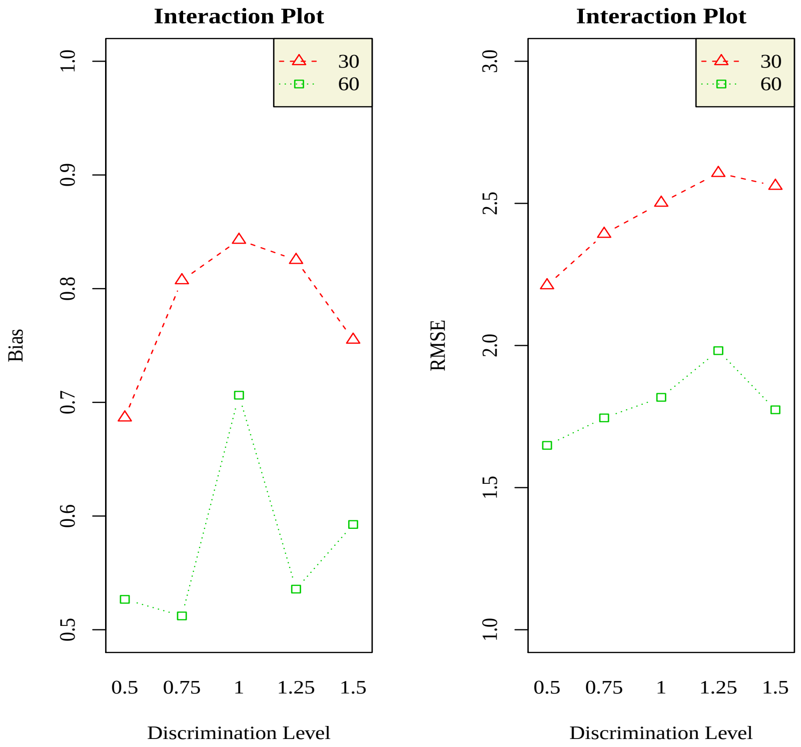


Figure 4.6: Bias and RMSE for b-parameter in the BF-C Model

4.2.2.3 Recovery of c-parameter

For pseudo-guessing parameter of this study, the effects of test length and discrimination level on bias and RMSE were examined using ANOVA procedure. Separate means and standard errors for pseudo-guessing are presented in Table A.3 of Appendix A. Both bias and RMSE were very small for c-parameter, that bias ranged from 0.10 and 0.04 and RMSE ranged from 0.0001 and 0.008, that this results demonstrated a very well recovery of psuedo-guessing parameter. Results from ANOVA in Table 4.4 shows the significant tests for test length and discrimination level on bias and RMSE. As c-parameter was very well estimated, there were non significant effects of the interaction between test length and discrimination level, $F(4,490) = 1.12, p = .290, \eta^2 < .001$, as well as main effects of the two studied factors on the bias, $F(1,490) = 27.20, p < .001, \eta^2 = .04$ and $F(4,490) = 34.42, p < .001, \eta^2 = .05$, respectively. The interaction between test length and discrimination level was not significant on bias, although the main effect of test length and main effect of discrimination level were both significant, however, the effect size were very small. For RMSE, all of the effects were not significant, $F(4,490) = 1.56, p = .210, \eta^2 = .002$ for the interaction, $F(1,490) = 1.560, p = .220, \eta^2 = .002$ for the effect of test length, and $F(4,490) = 1.10, p = .801, \eta^2 = .01$ for the effect of discrimination level of the specific ability dimensions. Figure 4.7 apparently shown these results in which there were very small bias and RMSE, thus, posited no different in c-parameter estimates to the true generated values across the studied test lengths and levels of discrimination.

Table 4.4: *Bias and RMSE for c-parameter in the BF-C Model*

Effect	*MS*	*F*	*df*	η^2
Bias				
Test Length *	<.001	27.20	1	.04
Discrimination Level*	<.001	34.42	4	.05
Test Length × Discrimination Level	<.001	1.12	4	<.001
Error	<.001		490	
RMSE				
Test Length	<.001	1.50	1	.002
Discrimination Level	<.001	1.10	4	.01
Test Length × Discrimination Level	<.001	1.56	4	.002
Error	<.001		490	

Note: RMSE: Root mean square error, MS: Mean square error, *df*: degrees of freedom, *p<.001.

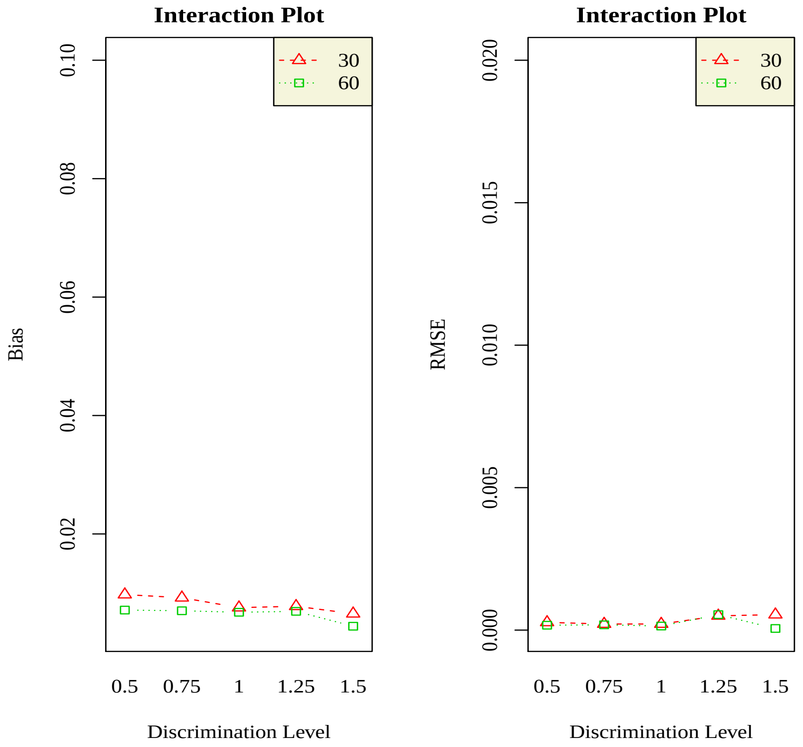


Figure 4.7: Bias and RMSE of c-parameter in BF-C Model

4.2.2.4 Recovery and Reliability of Thetas

Table A.4 in Appendix A summarizes the separate means and standard errors for bias and RMSE of all dimensions of thetas. For significant different tests of bias and RMSE of thetas, an RM-ANOVA was carried out and the results are shown in Table 4.5 and Table 4.6. The null hypothesis is that bias or RMSE of thetas does not change across different ability dimensions. The Mauchly's sphericity test showed a significant result that bias and RMSE of a-parameter do not meet the sphericity assumption, $W = .028, \chi(5) = 2466.01, p < .001$ and $W = .023$, $\chi(5) = 2592.74, p < .001$, respectively. These results suggesting that the observed matrices do not have approximately equal variances and equal covariances across ability dimensions. Thus, corrected RM-ANOVA F-tests are used such that Greenhouse-Geisser and Huynh-Feldt epsilon corrections were considered. The corrective coefficients were: Greenhouse-Geisser $\varepsilon_{bias} = .514$, $\varepsilon_{rmse} = .382$, and Huynh-Feldt $\varepsilon_{bias} = .513$, and $\varepsilon_{rmse} = .382$.

The multivariate approach for the within-runs was then used to evaluate the bias and RMSE of person parameter (thetas) as a function of ability dimension. The null hypothesis is that bias or RMSE of thetas does not change across different ability dimensions and with the effects of test length and discrimination level. Table Table 4.5 summarizes within-runs tests of the RM-MANOVA for the bias and RMSE of thetas. The Wilks' lambda (Λ) and Pillai's trace V are reported for each effect.

For the within-runs test, all effects were significant but there was no significant three-way interaction of ability dimension, test length and discrimination level for bias $F_{\Lambda}(12, 1291.40) = .640, p = .814$. These results show there were significant different of the bias of thetas from the effects of test length and discrimination level across ability dimensions. For RMSE, the three-way interaction of ability dimension, test length and discrimination level was significant, $F_{\Lambda}(12, 1291.40) = 1.83, p < .001$.

For between-runs significant tests, only the effect of test length was significant for the bias of thetas with small effect size, $F_{\Lambda}(1,490) = 135.61, p < .001, \eta^2 = .23$ and there was significant interaction between test length and discrimination level for the RMSE of thetas with a very small

Table 4.5: *Multivariate Within-runs Effects for Thetas in the BF-C Model*

Effect	Λ	V	$F_\Lambda\,(df_1,df_2)$	$F_V\,(df_1,df_2)$
Bias				
Dimension*	.003	.997	66782.3(3,488)	66782.3(3,488)
Dimension×TL*	.670	.300	98.43(3,488)	98.43(3,488)
Dimension×DL**	.970	.030	$1.096(12,1291.40)$	$1.94(12,1470)$
Dimension×TL×DL	.989	.011	$.64(12,1291.40)$	$.64(12,1470)$
RMSE				
Dimension*	.162	.838	1188.73(3,488)	1188.73(3,488)
Dimension×TL*	.705	.300	96.08(3,488)	96.08(3,488)
Dimension×DL*	.326	.680	$80.00(12,1291.40)$	$50.56(12,1470)$
Dimension×TL×DL*	.969	.031	$1.83(12,1291.40)$	$1.83(12,1470)$

Note: RMSE: Root mean square error, TL: Test Length, DL: Discrimination Level, df: degrees of freedom, *p<.001, **p<.01, Λ :Wilks' Lamdba, V : Pillai's trace.

effect size, $F_\Lambda\,(4,490) = 5.33, p < .001, \eta^2 = .02$. Table 4.6 shows the overall between-runs results for bias and RMSE of thetas, and Figure 4.8 illustrates the interaction between test length and discrimination level factors on bias and RMSE of thetas estimates.

Table 4.6: *Between-runs Effects for Thetas in the BF-C Model*

Effect	*MS*	*F*	df	η^2
Bias				
Test Length*	.012	135.61	1	.23
Discrimination Level	<.001	1.73	4	.002
Test Length × Discrimination Level	<.001	.53	4	.002
Error	<.001		490	
RMSE				
Test Length	<.001	0.000	1	.007
Discrimination Level*	.504	376.60	4	.24
Test Length × Discrimination Level**	.007	5.33	4	.02
Error	.001		490	

Note: RMSE: Root mean square error, MS: Mean square error, df: degrees of freedom, *p<.001, **p<.01.

In overall, the estimated thetas were all very well recovered as the bias was closed to zero with RMSE were closed to 1.0. There were very small different on the bias and and RMSE of thetas estimates between the 30-item and 60-item tests. Also, there was a slightly increased in the RMSE of primary theta ability when the discrimination level increases, and there was a slightly decreased in RMSE of the forth specific ability dimension when the discrimination level increases. In all other specific ability dimensions, the plots depict small different in bias and RMSE on the shorter 30-item test and the longer 60-item test as well as across five levels of discrimination in all ability dimensions.

In addition to the analysis of bias and RMSE of thetas estimates above, correlation of theta is also presented here to evaluate the accuracy of theta estimates because thetas in the primary or specific ability dimensions are important to evaluate subscore reliability and classification that will be presented in the next sections of this study. Pearson's correlations for all conditions between the generated and estimated subscores are summarized in Table A.4 in Appendix A and illustrated in Figure 4.9 below. These correlations were computed from equation presented in Chapter 3, that is the average of the observed squared correlations between the estimated subscores and the generated ability parameters over 50 replications.

The simulation correlation in all cases for all ability dimensions ranged from .67 and .82, where the correlation for the specific ability dimensions were higher than the primary dimension that is ranged from .88 and .98. This study observed an equal correlation in all ability dimensions for the 30-item test and 60-item test. The correlation of the ability estimates in the primary were lower than the correlation in the secondary ability dimensions. Also, a higher level of discrimination always has higher correlation in the primary ability dimension and lower levels of discrimination have slightly lower correlations. When the discrimination levels of the specific ability dimension were equal and higher than the primary ability dimension, the correlation of thetas were about equally high in all specific ability dimensions.

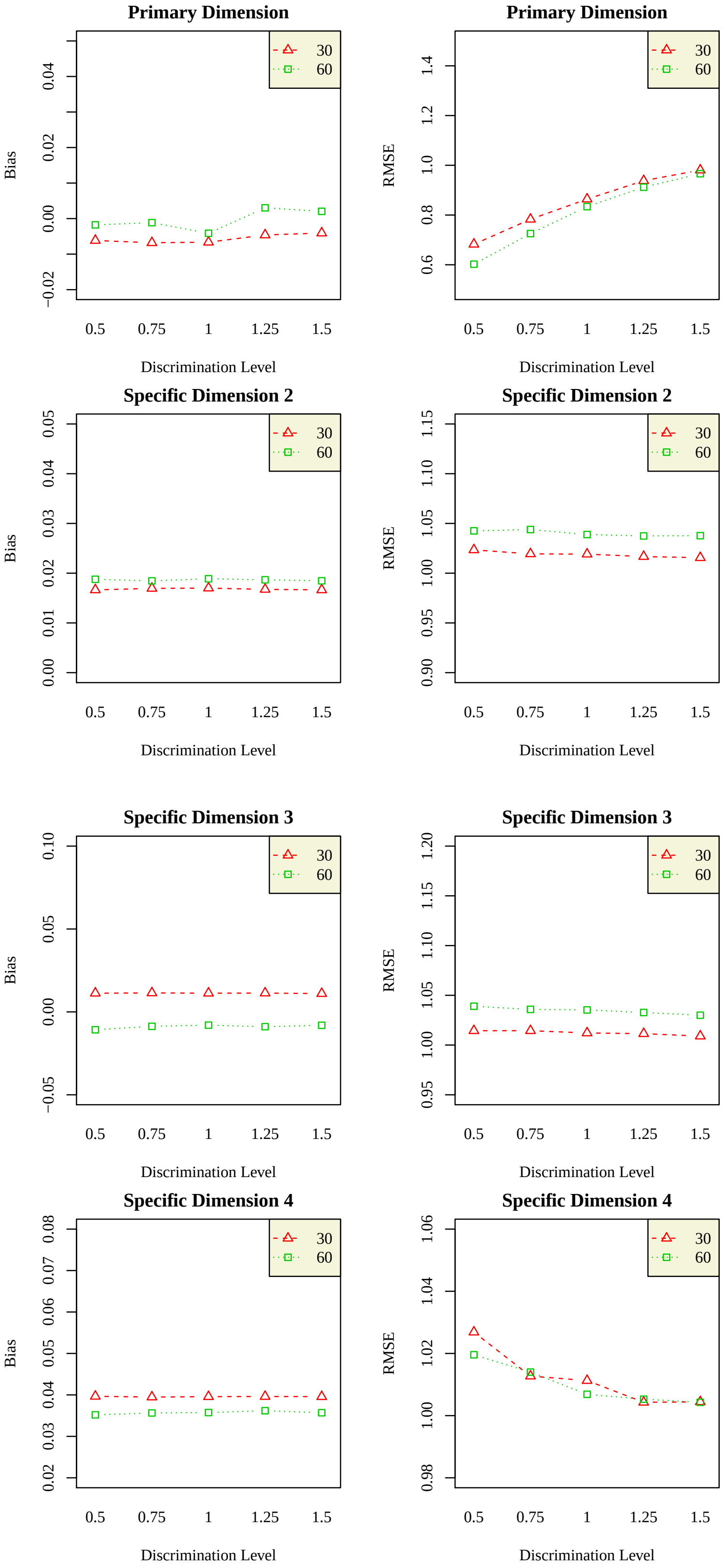


Figure 4.8: Interaction Plots between Test Length and Discrimination Level on Bias and RMSE for Each Theta in the BF-C Model

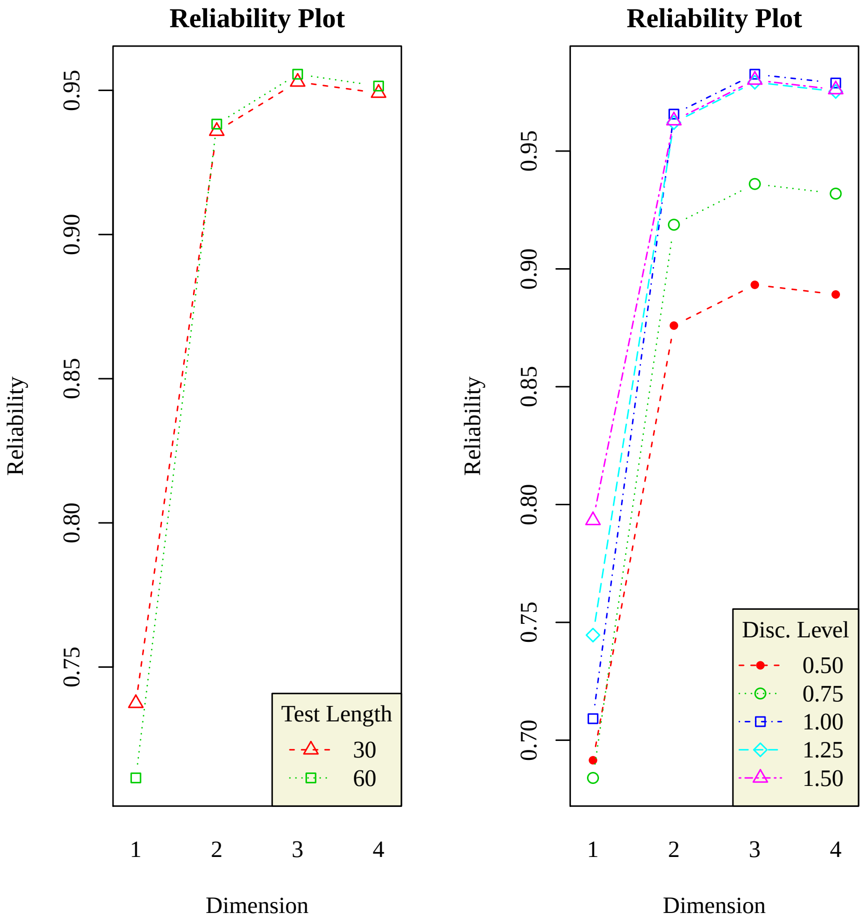


Figure 4.9: Reliability for the Estimated Subscores and Generated Abilities (Thetas)

4.2.3 Bayesian Criteria Comparison for Study 1

This section demonstrates comparisons of the Bayesian criteria for model complexity and fit at different test lengths and discrimination levels. The purpose of this section is solely to examine and demonstrate model-data fit across the two factors rather than for selecting or comparing the conditions in each or across factors.

Table 4.7 summarizes deviance, pD, DIC, AIC, BIC and -2 log Likelihood separated from different test lengths and discrimination levels. These Bayesian fit indices are graphically illustrated in Figure 4.10. The results show that there were stable changes within each studied factor in all of the Bayesian fit indices. In overall, Bayesian complexity and fit from Table 4.7 showed all the complexity of the 60-item test was as twice as big the 30-item test. This supports the MCMC estimations for a longer test to have more complex in terms of convergence of the MCMC that the estimation for all monitored parameters required more times than the shorter test. The distributions of the discrimination parameter that were varied in their discriminating levels between the primary and secondary ability dimensions were slightly influenced the MCMC convergence that as discrimination level increases the Bayesian complexity and fits tends to decrease with some fluctuations between the shorter and longer tests. Thus the distributions of the discrimination levels showed minor effects on the MCMC simulation convergence which in turn slightly induced the BF-C model Bayesian complexity and fit.

4.2.4 Subscore Reliability and Classification

4.2.4.1 Subscore Reliability

Results in the thetas recoveries section, in average, yield high accuracies in the estimated subscores to the true subscores of the specific ability dimensions (i.e. correlations ranged from .89 to .98). In this section, Bayesian marginal reliabilities, which is analogous to the coefficient alpha, are computed from Equation 3.5.1 for all estimated subscores using a program written in R. Subscore reliabilities based on the Bayesian marginal reliabilities for all ability dimension across the two test lengths and five levels of discrimination for BF-C model are summarized in Table 4.8. All subscores were observed to have very high reliabilities across all simulation conditions, that is ranged from .89 to .99.

Table 4.7: *BF-C Model Bayesian Complexity and Fit*

Test Length		Discrimination Level				
		.50	.75	1.00	1.25	1.50
30	deviance	53227.02	52648.18	51560.66	49188.47	47701.54
	pD	13204.17	10945.36	10959.02	10100.01	9514.12
	DIC	66431.19	63593.54	62519.68	59288.49	57215.66
	AIC	79635.36	74538.9	73478.7	69388.5	66729.78
	BIC	149792.04	132694.02	131706.39	123052.1	117280.38
	-2 Log L	41177.05	42328.53	41542.07	39893.32	39573.48
60	deviance	108385.76	105102.95	105310.42	97548.81	98623.7
	pD	9166.38	9287.48	7696.22	7982.14	6696.21
	DIC	117552.14	114390.43	113006.65	105530.95	105319.9
	AIC	126718.52	123677.91	120702.87	113513.09	112016.11
	BIC	175421.51	173024.35	161594.61	155923.94	147594.53
	-2 Log L	99713.09	96495.34	97951.33	90231.33	92249.37

Table 4.8: *Subscore Bayesian Marginal Reliability from BF-C Model*

Test Length	Discrimination Level				
Dimension	.50	.75	1.00	.125	1.50
30					
1	.97	.96	.96	.96	.98
2	.97	.97	.98	.97	.97
3	.92	.95	.97	.97	.97
4	.89	.89	.93	.93	.97
60					
1	.98	.98	.98	.98	.99
2	.97	.97	.98	.97	.99
3	.98	.98	.97	.98	.99
4	.99	.97	.93	.98	.98

Figure 4.11 illustrates plots of Bayesian marginal reliabilities for all subscores across test length and discrimination levels. These plots showed higher subscore reliabilities in the 60-item test compared to the 30-item test, and increasing reliabilities as a function of discrimination level. There is a slightly decreased of this high subscore reliability in the 60-item test when the discrimination level of the specific ability dimension has an equal level of discrimination (i.e. 1.00) to the primary ability dimension. The highest level of discrimination yielded the highest subscore reliability. The last plot in Figure 4.11 showed equally high reliabilities between the primary ability dimension and specific ability dimensions. Only the forth dimension showed slightly a lower high subscore reliability when the discrimination levels were smaller than 1.50.

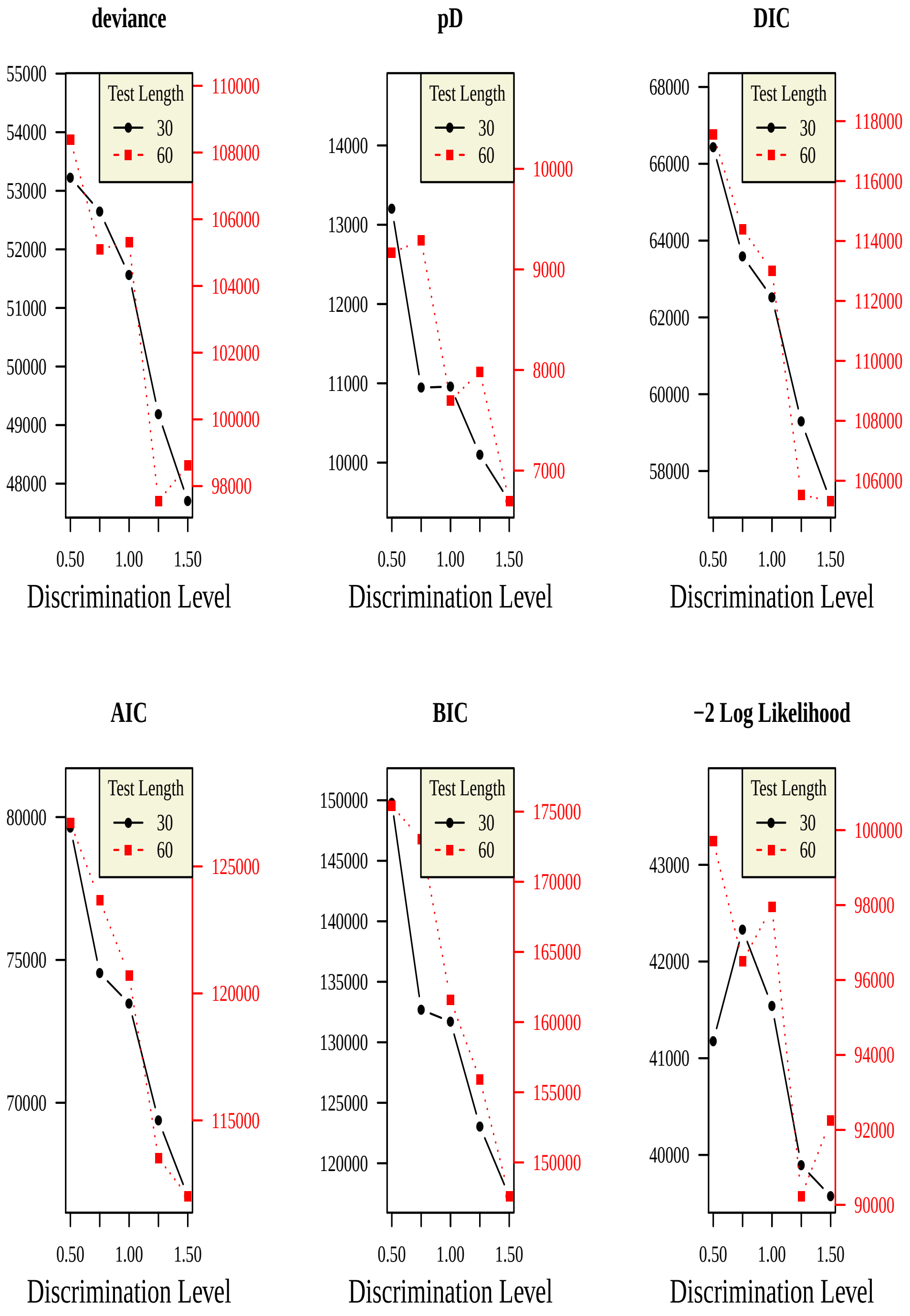


Figure 4.10: Bayesian Indices for BF-C Model Complexity and Fit

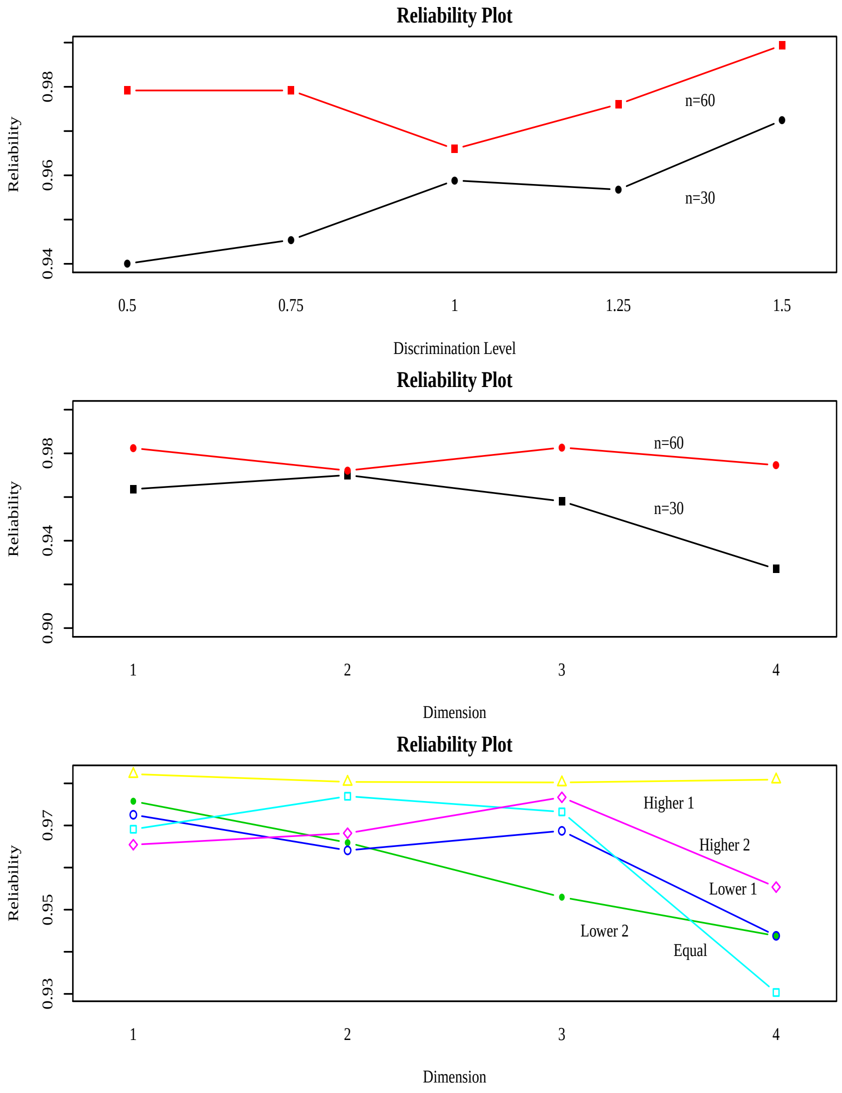


Figure 4.11: Subscore Reliability for BF-C Model

4.2.4.2 Subscore Separation Index

To investigate further, subscore separation index (SSI) was computed to evaluate the quality of the subscores from the specific ability dimension that may be used for test reporting purposes. Table 4.9 summarizes the SSI for each pair of primary and specific ability dimension, which is also illustrated in Figure 4.12. The 30-item test showed lower SSIs than test with 60 items, and SSI goes up when discrimination level increases. The SSI was ranged between .066 and .083 for the 30-item test and ranged between .123 and .147 for the 60-item test. The associated hit rate for SSI is linearly related and it is presented as a percentage of examinees have SSI greater than 1.0.

This model observed that, as discrimination level increases, about less than 10% of the examinees demonstrated SSI greater than 1.00 in the 30-item test and the hit rates were twice larger for the 60-item test. There were increasing patterns of SSI and hit rate when the discrimination levels in the specific ability dimension increases. These results infer higher SSI and hit rate for a longer test is administered. As the level of discrimination of the specific ability dimension is higher than the primary ability, the SSI and hit rate tend to increase. More about these results are discussed in Chapter 5.

Table 4.9: *Subscores Separation Index Means, Standard Error and Hit Rate for BF-C Model*

	Discrimination Level														
Test Length	.50			.75			1.00			1.25			1.50		
	1-2	1-3	1-4	1-2	1-3	1-4	1-2	1-3	1-4	1-2	1-3	1-4	1-2	1-3	1-4
SSI															
30 Mean	.066	.068	.068	.071	.071	.070	.078	.077	.078	.079	.080	.078	.084	.085	.083
s.e	.025	.027	.025	.028	.028	.033	.032	.032	.033	.033	.031	.031	.036	.037	.034
60 Mean	.123	.124	.122	.132	.131	.131	.129	.133	.130	.133	.134	.132	.147	.145	.143
s.e	.038	.038	.046	.046	.041	.040	.041	.042	.039	.049	.047	.044	.047	.045	.045
Hit Rate															
30 (%)	6.6	6.8	6.8	7.1	7.1	7.0	7.8	7.7	7.8	7.9	8.0	7.8	8.4	8.5	8.3
60 (%)	12.3	12.4	12.2	13.2	13.1	13.1	12.9	13.3	13.0	13.3	13.4	13.2	14.7	14.5	14.3

Note: SSI: Subscore Separation Index

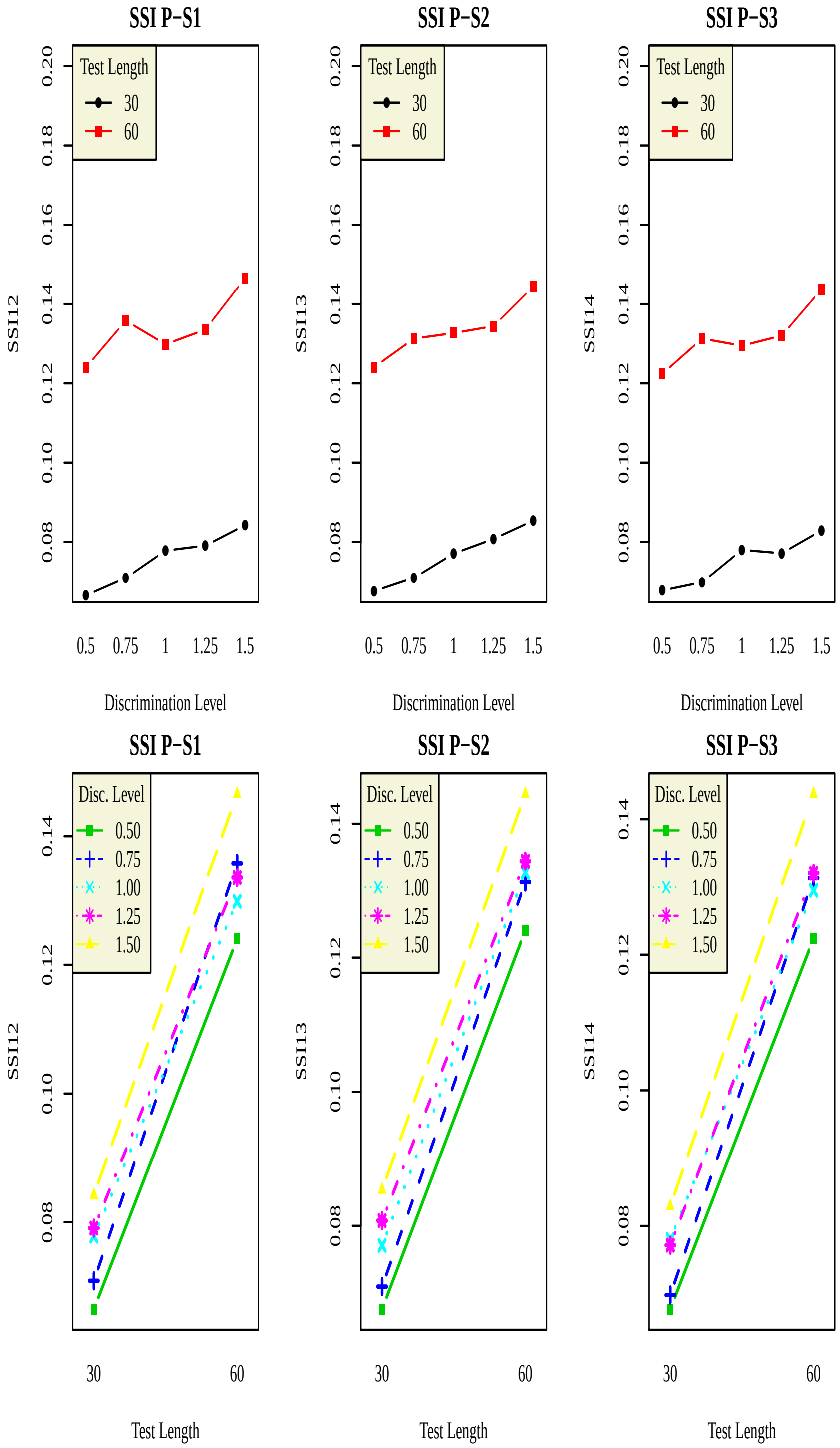


Figure 4.12: Plot of SSI and Hit Rate for BF-C Model

4.3 Results For Study 2

For this study, sets of true data were generated from the bi-factor partially compensatory MIRT model (BF-PC), specifically the M3PL model discussed in Equation 3.1.3 to Equation 3.1.4 in Chapter 3.

4.3.1 MCMC Simulation Diagnostics for Study 2

For Bayesian convergence criterion in Study 2, Figure 4.13 and Figure 4.14 illustrate the plots of $\hat{R}$ for item parameters on 30-item test and 60-item test at five levels of item discrimination. Figure 4.15 and Figure 4.16 show the plots of $\hat{R}$ for the examinee parameters. For all 50 replications, The plots of $\hat{R}$ for both item and examinee parameters for this study showed $\hat{R}$ less than 1.20 that the output from all parallel chains of the MCMC simulations is indistinguishable, and all of the monitored item and examinee parameters have eventually converged to the stationary distribution which represents the joint posterior distribution as in the Equation 3.2.5 in Chapter 3. The trace plots for all selected parameters in Figure B.1 to Figure B.4 show good mixing as the draws from MCMC simulation from the two chains were moving around the same parameter space of the statistical distributions. Also, the curves from the Kernel density plots in each selected monitored parameter also show smooth curves that their posteriors come from the desired posterior distributions.

4.3.2 Parameter Recovery for Study 2

For this model, SAS PROC GLM procedure for RM-ANOVA was used to evaluate bias and RMSE for item discriminations, item difficulties, and examinee abilities (i.e. thetas) that have been estimated at four different ability dimensions. ANOVA was used for evaluating the recovery of pseudo-guessing parameter. The same between-subjects factors defined in Study 1, that is test lengths and discrimination levels, were examined in this study.

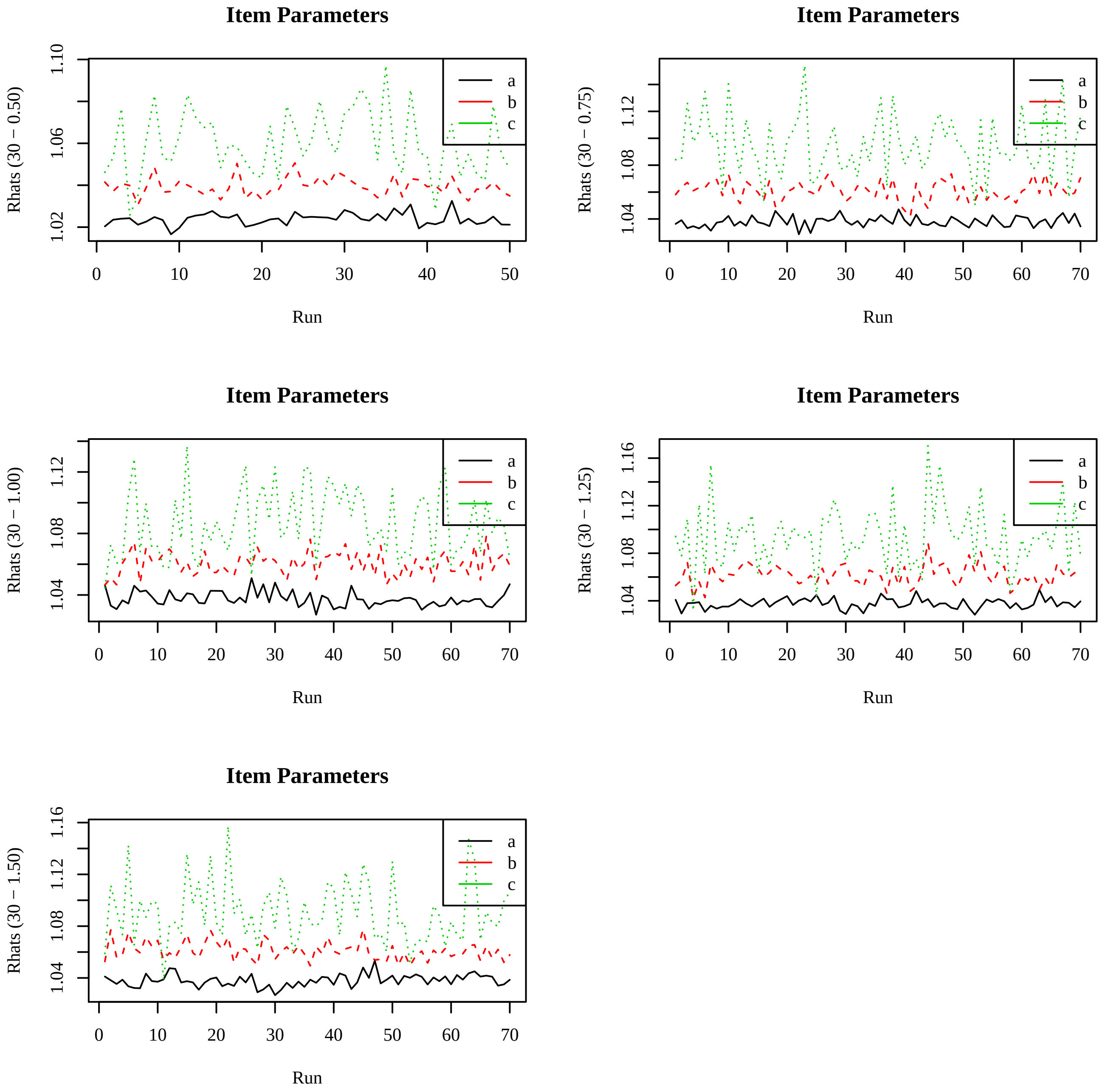


Figure 4.13: Rhats for Item Parameters in the BF-PC Model for 30-item Test

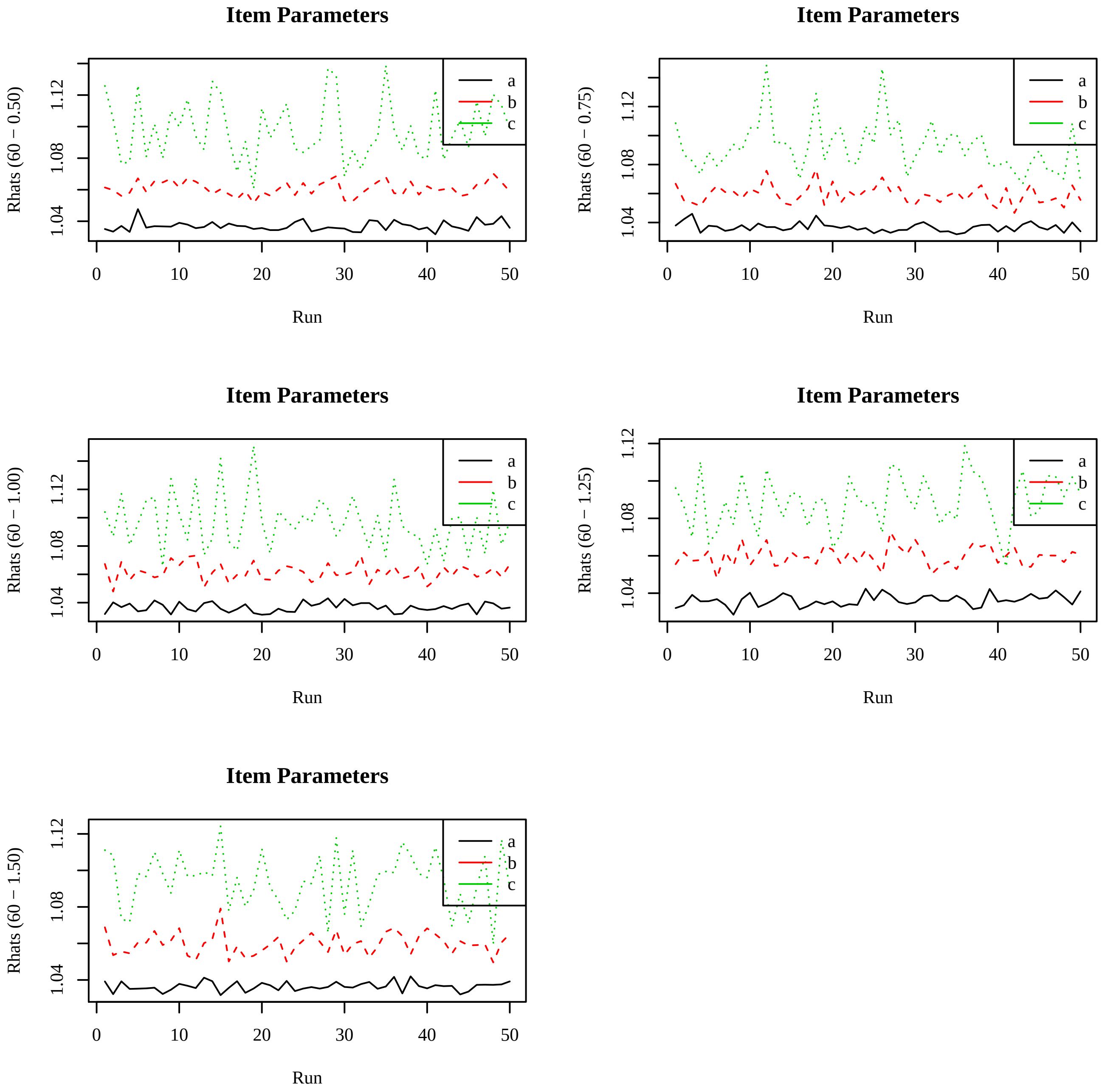


Figure 4.14: Rhats for Item Parameters in the BF-PC Model for 60-item Test

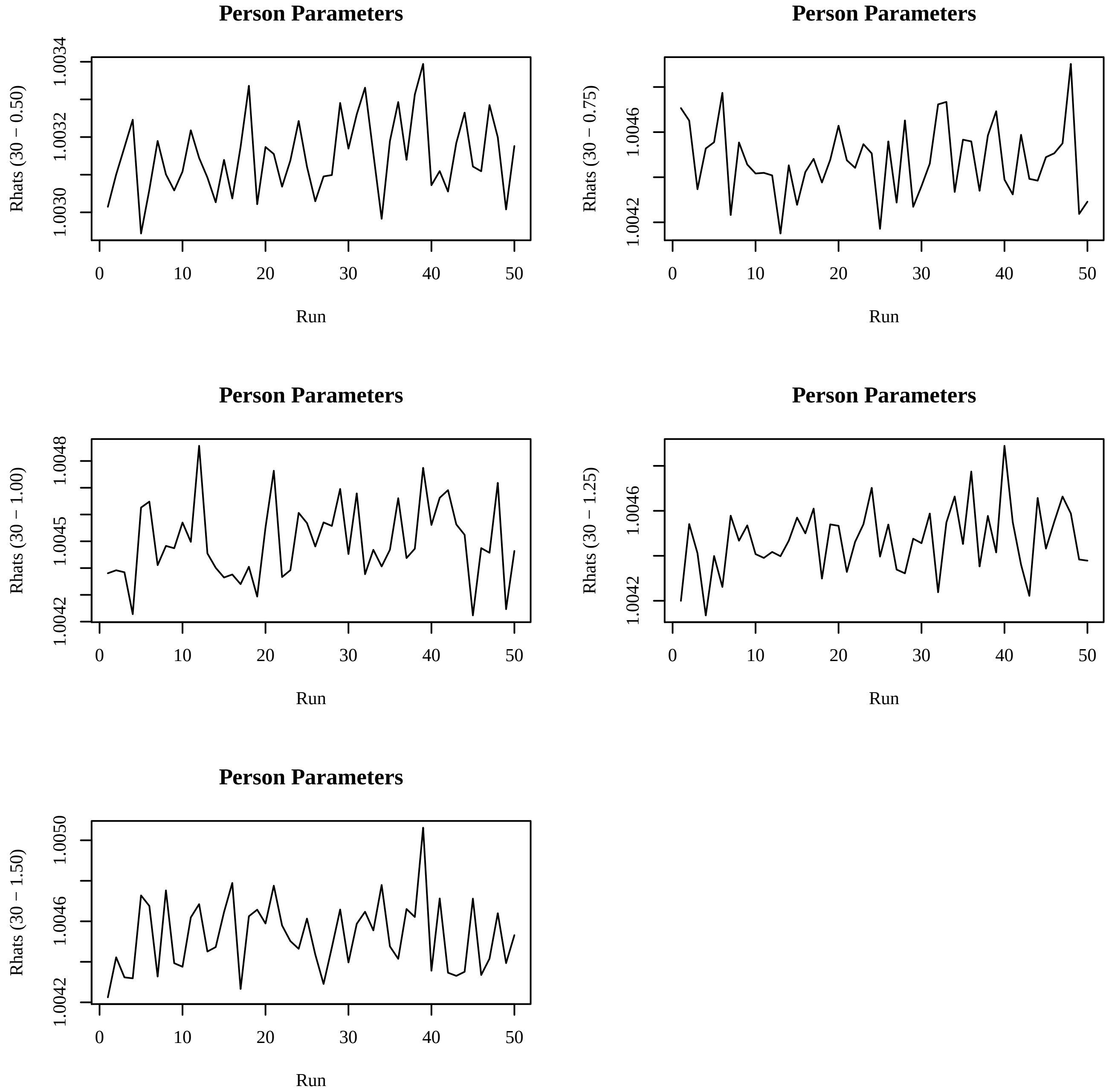


Figure 4.15: Rhats for Examinee (or Person) Parameters in the BF-PC Model for 30-item Test

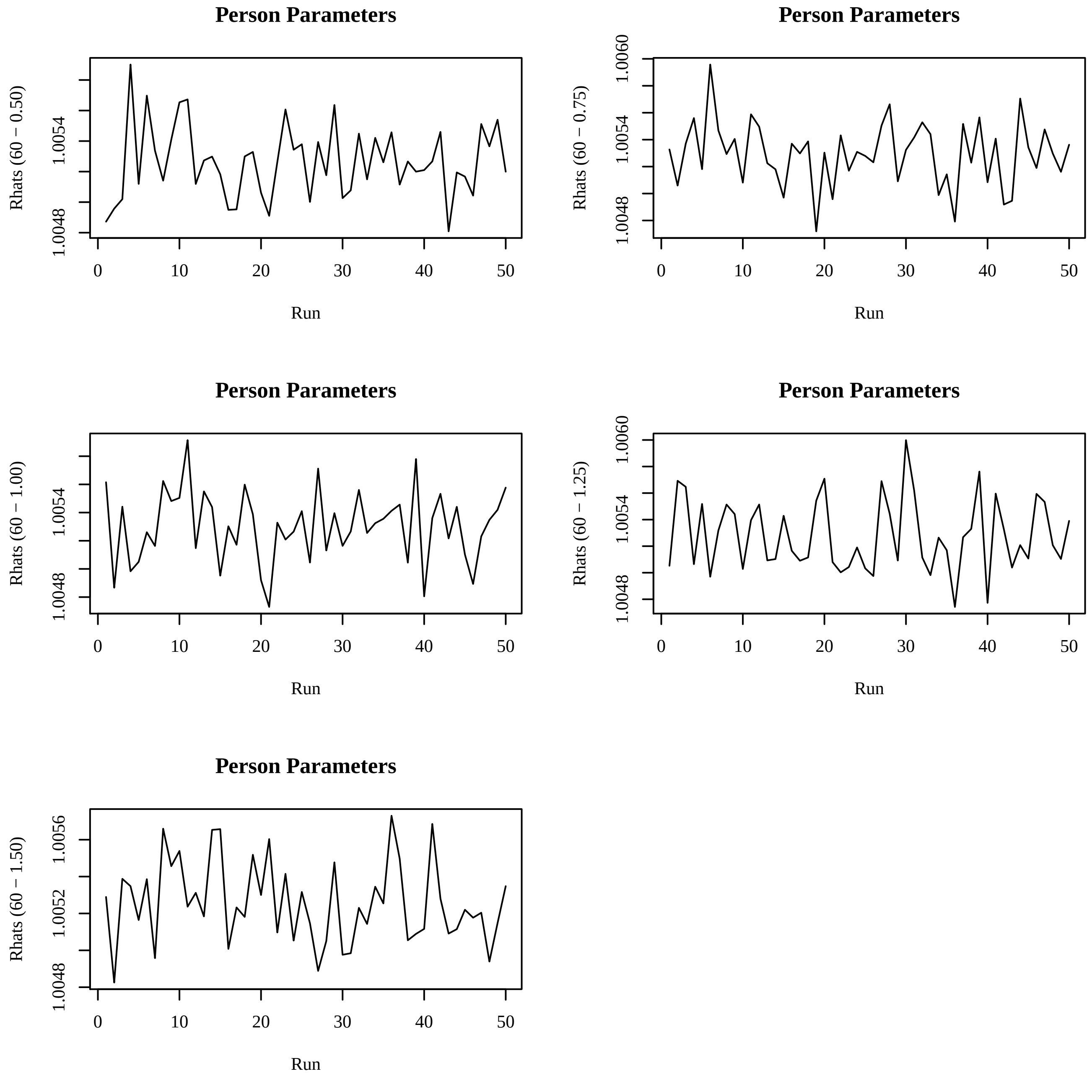


Figure 4.16: Rhats for Examinee (or Person) Parameters in the BF-PC Model for 60-item Test

4.3.2.1 Recovery of a-parameter

From Table B.1 in Appendix B it follows that bias of the a-parameters was positive with decreasing bias as the level of discrimination increases. For the 30-item test, the estimated a-parameters were equally biased compared to the 60-item test across five discrimination levels for the secondary factors dimension. Also, the a-parameters in the primary dimension showed more biased than the a-parameters in the specific ability dimensions. Finally, the RMSE for a-parameter in this model decreases at higher discrimination levels and there were no clear distinctions in RMSEs between the two test lengths.

For the RM-ANOVA analysis, the Mauchly's sphericity test showed a significant result that bias and RMSE of a-parameter do not meet the sphericity assumption, $W = .818, \chi(5) = 98.03, p < .001$and $W = .967, \chi(5) = 16.50, p < .01$ respectively. These reseults suggesting that the observed matrices do not have approximately equal variances and equal covariances. Thus, corrected RM-ANOVA F-test is used such that Greenhouse-Geisser and Huynh-Feldt epsilon corrections were considered. The corrective coefficient were: Greenhouse-Geisser $\varepsilon_{bias} = .876$, $\varepsilon_{rmse} = .980$, and Huynh-Feldt $\varepsilon_{bias} = .881$, and $\varepsilon_{rmse} = .986$.

The multivariate approach for the within-runs tests was then used to evaluate the bias and RMSE of a-parameter as a function of ability dimension. The null hypothesis is that bias or RMSE of a-parameter does not change across different ability dimensions. Table 4.10 and Table 4.11 summarize RM-MANOVA for bias and RMSE for a-parameter. The Wilks' lambda (Λ) and Pillai's trace (V) are reported for each effect. All two-way interactions and main effects were significant. These results suggest that test lengths and discrimination levels, respectively, interacts with ability dimensions, e.g. $F_{\Lambda}(3,488) = 9.91, p < .001$ and $F_{\Lambda}(12,1291.40) = 182.81, p < .001$ for bias, and $F_{\Lambda}(3,488) = 11.71, p < .001$ and $F_{\Lambda}(12,1291.40) = 274.80, p < .001$ for RMSE of a-parameter.

This study observed that bias and RMSE of a-parameter across different ability dimensions depends upon test length and discrimination level. However, the three-way interaction of ability dimension, test length and discrimination level for both bias and RMSE were not significant, $F_{\Lambda}(12,1291.40) = 1.14, p = .326$ and $F_{\Lambda}(3,488) = 1.73, p = .086$. In overall, bias and

RMSE of a-parameter were substantially significantly change with ability dimension, $F_{\Lambda}(3,488) = 21686.3, p < .001$and $F_{\Lambda}(3,488) = 8807.27, p < .001$.

For the between-runs effects, the interaction between test length and discrimination level was not significant for the bias of a-parameter, $F(4,490) = 2.21, p = .08$, and significant for the RMSE of a-parameter, $F(4,490) = 2.76, p < .05$. Both main effects of test length and discrimination level were significant, for bias $F(1,490) = 23.52, p < .001, \eta^2 = .09$ and $F(4,490) = 4821.67, p < .001, \eta^2 = .97$, and RMSE $F(1,490) = 7.90, p < .001, \eta^2 = .08$ and $F(4,490) = 6605.56, p < .001, \eta^2 = .98$.

Although there was a statistical significant different of the bias and RMSE of a-parameter between the 30-item and 60-item tests, the effect sizes were very small. There were a substantially large effect sizes on the significant different across the discrimination levels of the bias and RMSE of a-parameter. Figure 4.17a and 4.17b illustrate these effects. The plots showed decreasing bias and RMSE of the a-parameter from primary ability dimension to secondary ability dimensions, and there was very similar bias and RMSE of the a-parameter between the two test lengths.

Also, bias and RMSE of the a-parameter decrease when the level of discrimination increases. This gives the ability dimension with the highest discrimination has the lesser biased in the a-parameter. It also showed that only RMSE of the a-parameter at the smallest discrimination has unchanged across ability dimensions (see red dotted line in the interaction plot of RMSE for a-parameter at different discrimination levels in Figure 4.17b.)

Table 4.10: *Multivariate Within-runs Effects for a-parameter in the BF-PC Model*

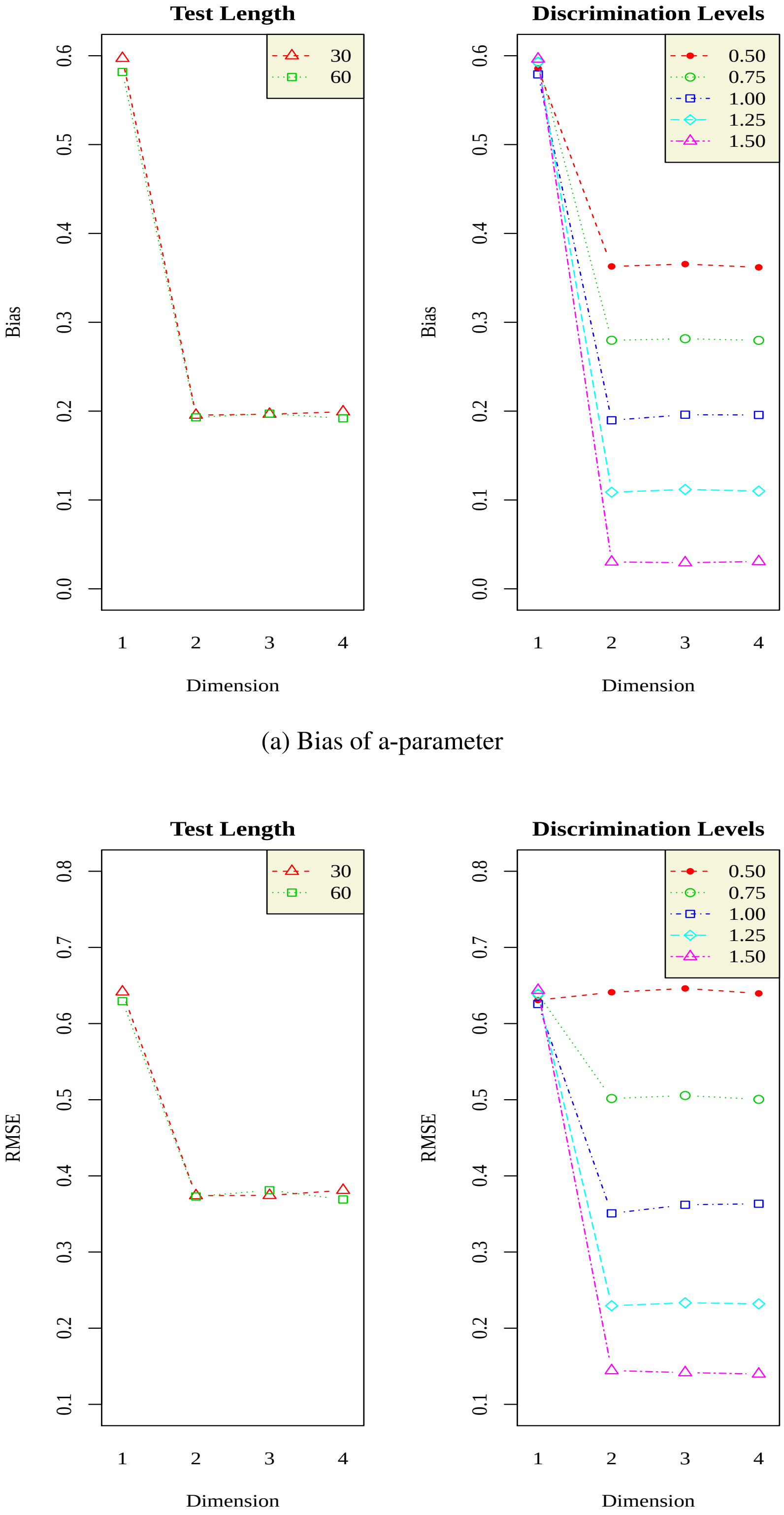
Effect	Λ	V	$F_{\Lambda}\,(df1,df2)$	$F_{V}\,(df1,df2)$
Bias				
Dimension*	.007	.993	$21686.3(3,488)$	$21686.3(3,488)$
Dimension×TL*	.943	.057	$9.91(3,488)$	$9.91(3,488)$
Dimension×DL*	.073	.935	$182.81(12,1291.40)$	$55.44(12,1470)$
Dimension×TL×DL	.973	.028	$1.14(12,1291.40)$	$1.14(12,1470)$
RMSE				
Dimension*	.018	.982	$8807.27(3,488)$	$8807.27(3,488)$
Dimension×TL*	.933	.067	$11.71(3,488)$	$11.71(3,488)$
Dimension×DL*	.035	,.980	$274.80(12,1291.40)$	$59.44(12,1291.40)$
Dimension×TL×DL	.959	.042	$1.73(12,1291.40)$	$1.73(12,1291.40)$

Note: RMSE: Root mean square error, TL: Test Length, DL: Discrimination Level, df: degrees of freedom, *p<.001, **p<.01, Λ :Wilks' Lamdba, V : Pillai's trace.

Table 4.11: *Between-runs Effects for a-parameter in the BF-PC Model*

Effect	MS	F	df	η^2
Bias				
Test Length*	.019	23.52	1	.09
Discrimination Level*	3.873	4821.67	4	.97
Test Length × Discrimination Level	.007	2.21	4	
Error	.0008		490	
RMSE				
Test Length*	.011	7.90	1	.08
Discrimination Level*	9.044	6605.56	4	.98
Test Length × Discrimination Level	.004	2.76	4	
Error	.671		490	

Note: RMSE: Root mean square error, MS: Mean square error, df: degrees of freedom, *p<.001.



(a) Bias of a-parameter

(b) RMSE of a-parameter

Figure 4.17: Interaction Plots of Test Length, Discrimination Level and Dimension on Bias and RMSE for a-parameter in BF-PC Model

4.3.2.2 Recovery of b-parameter

The same procedure of RM-ANOVA was carried out to examine the recovery of b-parameter. Table B.2 in Appendix B shows separate means and standard errors for bias and RMSE of b-parameter for this model. In average, the estimates of b-parameter were not very well recovered that the bias and RMSE ranged from .50 to 1.70.

The Mauchly's sphericity test showed a significant result that bias and RMSE of a-parameter do not meet the sphericity assumption, $W = .944, \chi(5) = 27.74, p < .001$ and $W = .940, \chi(5) = 30.15, p < .001$ respectively. These reseults suggesting that the observed matrices do not have approximately equal variances and equal covariances. Thus, corrected RM-ANOVA F-test is used such that Greenhouse-Geisser and Huynh-Feldt epsilon corrections were considered. The corrective coefficient were: Greenhouse-Geisser $\varepsilon_{bias} = .963$, $\varepsilon_{rmse} = .964$, and Huynh-Feldt $\varepsilon_{bias} = .980$, and $\varepsilon_{rmse} = .970$.

The null hypothesis is that bias or RMSE of b-parameter does not change across different ability dimensions. Table 4.12 and Table 4.13 summarize RM-MANOVA for bias and RMSE for b-parameter. The Wilks' lambda (Λ) and Pillai's trace (V) are reported for each effect.

In the within-runs effects, only interactions between dimensions and test length were significant, $F_{\Lambda bias}(3,488) = 23.72, p < .001$ and $F_{\Lambda rmse}(3,488) = 11.30, p < .001$ respectively. Main effects of dimension on the bias and RMSE of b-parameter were significant, $F_{\Lambda bias}(3,488) = 7966.77$, $p < .001$ and $F_{\Lambda rmse}(3,488) = 170.96$,$p < .001$. There were no interactions between ability dimension and discrimination level in both bias and RMSE, $F_{\Lambda bias}(12,1291.40) = .007$, $p = 1.00$ and $F_{\Lambda rmse}(12,1291.40) = .65$,$p = .802$, as well as the three-way interaction effects, $F_{\Lambda bias}(12,1291.40) = .04, p = 1.00$ and $F_{\Lambda rmse}(12,1291.40) = .34, p = .982$, respectively. Thus, bias and RMSE of b-parameter significantly change with ability dimension, and test length interacts across ability dimensions.

For the between-runs effects, none of the interactions and main effects were significant for bias and RMSE of b-parameter. This is supported by a very small effect size on this non significant result, for example there was a very small (non significant) effect of test length, $\eta^2_{bias} = .02$

and $\eta^2_{rmse} = .01$ respectively. Figure 4.18 clearly illustrates these results, and in all cases, bias and RMSE for the primary ability dimension were somewhat larger than bias and RMSE in the secondary ability dimensions. The plots showed unnoticeable different in bias and RMSE of b-parameter when a test has 30 items or 60 items. The plots also showed the same conclusion that there were no clear different in bias and RMSE of b-parameter across different levels of discrimination.

Table 4.12: *Multivariate Within-runs Effects for b-parameter in the BF-PC Model*

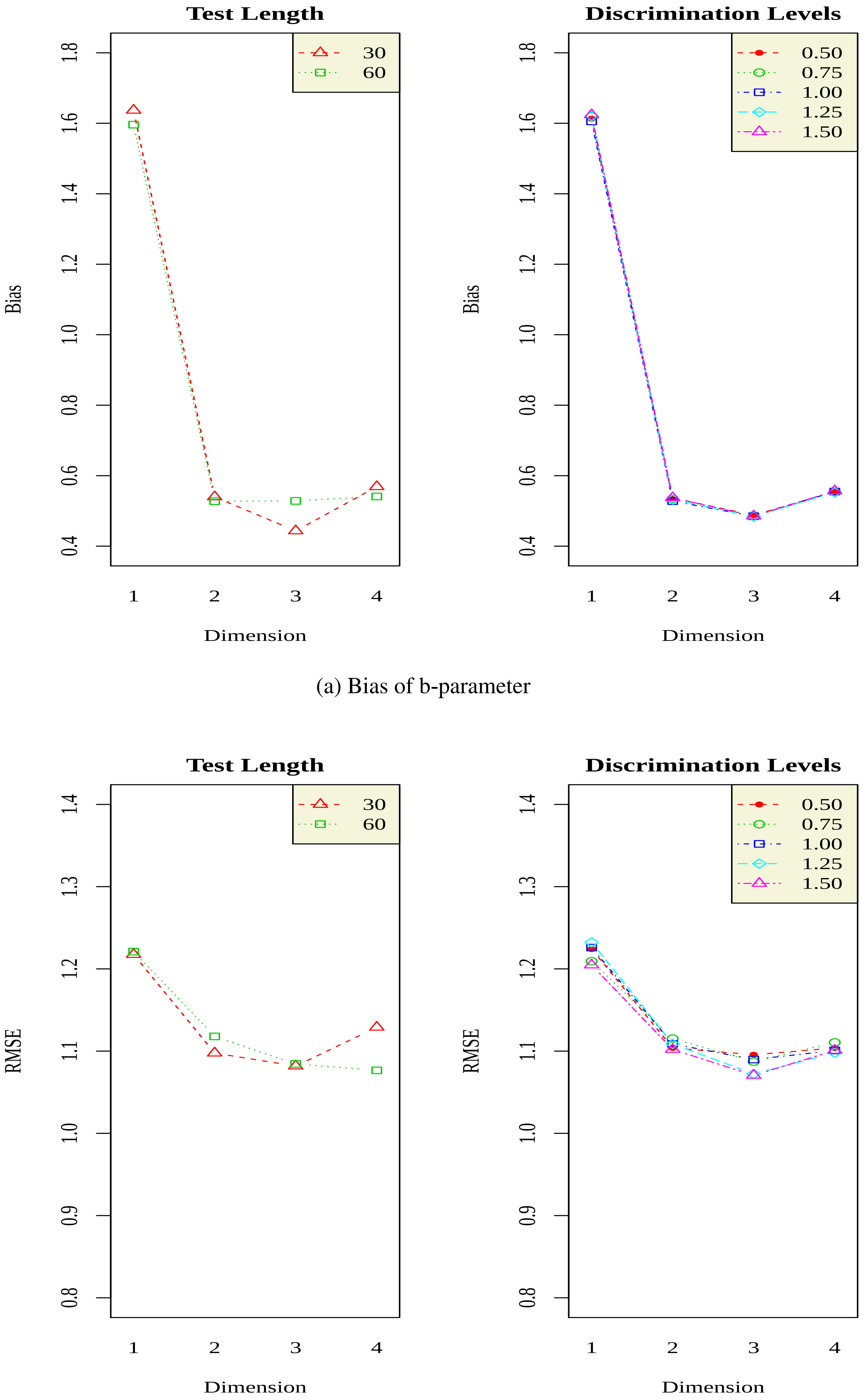
Effect	Λ	V	$F_\Lambda\,(df_1,df_2)$	$F_V\,(df_1,df_2)$
Bias				
Dimension*	.020	.980	7966.77(3,488)	7966.77(3,488)
Dimension×TL*	.873	.127	23.72(3,488)	23.72(3,488)
Dimension×DL	.998	.002	.007(12,1291.40)	.07(12,1470)
Dimension×TL×DL	.999	.001	.04(12,1291.40)	.04(12,1470)
RMSE				
Dimension*	.488	.512	170.96(3,488)	170.96(3,488)
Dimension×TL*	.935	.065	11.30(3,488)	11.30(3,488)
Dimension×DL	.984	.016	.65(12,1291.40)	.65(12,1470)
Dimension×TL×DL	.992	.008	.34(12,1291.40)	.34(12,1470)

Note: RMSE: Root mean square error, TL: Test Length, DL: Discrimination Level, df: degrees of freedom, *p<.001.

Table 4.13: *Between-runs Effects for b-parameter in the BF-PC Model*

Effect	MS	F	df	η^2
Bias				
Test Length	.001	.02	1	.02
Discrimination Level	.003	.13	4	.0001
Test Length × Discrimination Level	.002	.08	4	
Error	.024		490	
RMSE				
Test Length	.021	1.60	1	.01
Discrimination Level	.010	.78	4	.006
Test Length × Discrimination Level	.013	.96	4	.001
Error	.013		490	

Note: RMSE: Root mean square error, MS: Mean square error, df: degrees of freedom.



(a) Bias of b-parameter

(b) RMSE of b-parameter

Figure 4.18: Interaction Plots of Test Length, Discrimination Level and Dimension on Bias and RMSE for b-parameter in BF-PC Model

4.3.2.3 Recovery of c-parameter

For pseudo-guessing parameter in BF-PC model, the effects of test length and discrimination level on bias and RMSE were examined using ANOVA procedure. The separate means and standard errors for pseudo-guessing are presented in Table B.3 of Appendix B. On average, bias for c-parameter was .12 and RMSE was .015.

The results from ANOVA in Table 4.14 showed all of the effects were not significant (all $p >$ .50 and $p >$ .60 for bias and RMSE, respectively). Thus, bias and RMSE for c-parameter did not change significantly as test length or discrimination level increases (or decreases). Plots in Figure 4.19 illustrate bias and RMSE for 30-item and 60-item tests that appear very closed to each other across five levels of discrimination.

Table 4.14: *ANOVA for Bias and RMSE of c-parameter in BF-PC Model*

Effect	*MS*	*F*	*df*	η^2
Bias				
Test Length	<.0001	.02	1	.001
Discrimination Level	<.0001	.13	4	.003
Test Length × Discrimination Level	<.0001	.08	4	
Error	.008		490	
RMSE				
Test Length	.021	1.60	1	.01
Discrimination Level	.010	.78	4	.006
Test Length × Discrimination Level			4	
Error			490	

Note: RMSE: Root mean square error, MS: Mean square error, *df*: degrees of freedom.

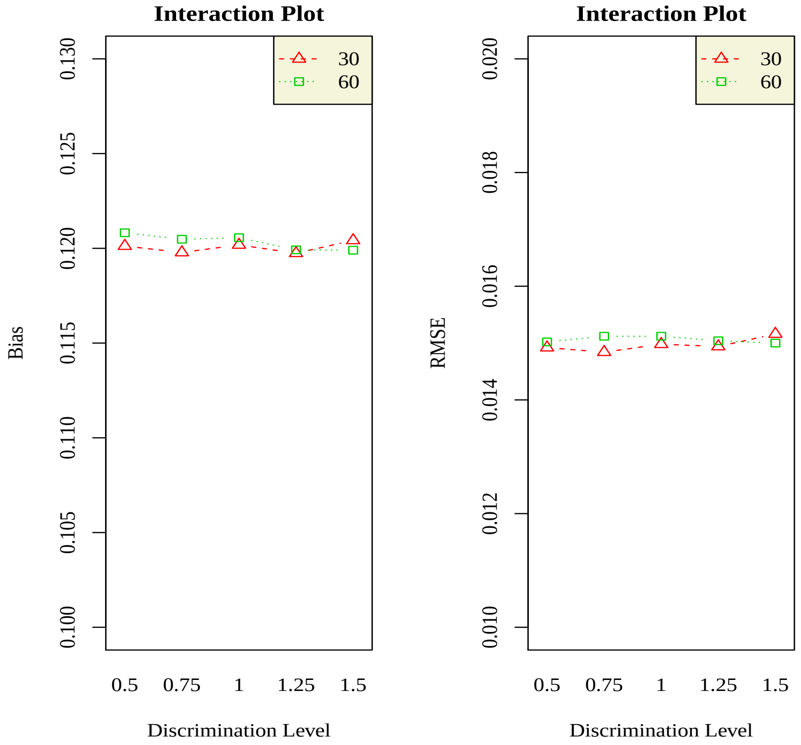


Figure 4.19: Interaction Plots of Test Length and Discrimination Level on Bias and RMSE for c-parameter in BF-PC Model

4.3.2.4 Recovery and Reliability of Thetas

Table B.4 in Appendix B contains bias, RMSE and Pearson's correlation for abilities (thetas). For all conditions, thetas were all well recovered as the biases were closed to zero and RMSEs, in average, were closed to 1.0. Using SAS PROC GLM procedure, a series of ANOVA was necessary instead of RM-ANOVA for these uncorrelated (nonsignificant) thetas. ANOVAs were carried out to examine the effects of test length and discrimination levels on bias and RMSE on each dimension of latent ability theta. Table 4.15 shows the results from ANOVA on bias and RMSE. All of the interactions between test length and discrimination level were not significant indicating that there were no interaction between these two factors on the bias and RMSE of theta in each dimension. This is supported by very small effect sizes which ranged from .01 to .07 for biases and from .01 to .08 for RMSEs.

The main effects of test length and discrimination level were significant on bias and RMSE of thetas for all dimensions. There was large effect size of test length on the bias for the primary dimension $\left(\eta_1^2 = .82\right)$, and medium effect sizes of biases on thetas in the secondary dimensions $\left(\eta_2^2 = .36, \eta_3^2 = .34, \eta_4^2 = .57\right)$. Also, there were small effect sizes of discrimination level for all biases on thetas $\left(\eta_1^2 = .02, \eta_2^2 = .02, \eta_3^2 = .03, \eta_4^2 = .13\right)$. Whereas for RMSE, there was a large effect of test length on the primary ability dimension $\left(\eta_1^2 = .65\right)$ and small effects on the secondary ability dimensions, $\left(\eta_2^2 = .22, \eta_3^2 = .03, \eta_4^2 = .15\right)$. Conversely, there was small effect size of discrimination level on the primary ability dimension $\left(\eta_1^2 = .02\right)$ but slightly large effect sizes on the secondary ability dimensions, $\left(\eta_2^2 = .41, \eta_3^2 = .49, \eta_4^2 = .47\right)$.

Figure 4.20 demonstrates the effects of test length and discrimination level on bias and RMSE for theta in each dimension. The left hand side of the plot shows that for each ability dimension the bias of theta was not much changed from the lowest to the highest level of discrimination. Most of all of the estimated thetas were negatively biased towards zero.

In overall, at all discrimination levels, the bias of theta for 30-item test conditions in the primary, specific dimension 3 and specific dimension 4 showed less bias compared to the 60-item test condition. Although an opposite result was observed for the specific dimension 2, however, all the changes of the biases were on a small scale that is closed to zero.

It is expected that all RMSE for thetas were closed to 1.0. All of the 60-item test conditions were observed to have lower RMSEs than the 30-item test conditions. There were slight decreased in the RMSE as the discrimination level increases for the specific ability dimensions, and RMSE was unchanged for the primary dimension when discrimination level goes up.

Correlation of theta is also presented here to evaluate the accuracy of theta estimates because thetas in the primary or specific ability dimensions are important to evaluate subscore reliability and classification. Pearson's correlations for all conditions between the generated and estimated subscores are also summarized in Table B.4 of Appendix B, and the plots are illustrated in Figure 4.21.

These estimated correlations were computed from equation presented in Chapter 3, that is the average of the observed squared correlations between subscores and the generated ability pa-

Table 4.15: *ANOVA for Bias and RMSE of Theta in Each Dimension from the BF-PC Model*

Effect	*MS*				*F*				df	η^2			
Dimension	1	2	3	4	1	2	3	4		1	2	3	4
Bias													
Test Length*	+	+	+	+	2189.88	274.39	262.31	655.73	1	.82	.36	.34	.57
Discrimination Level*	+	+	+	+	2.40	2.80	3.43	1.68	4	.02	.02	.03	.13
TL × DL	+	+	+	+	0.70	1.40	2.31	2.66	4	.01	.01	.02	.01
Error	†	†	†	†					490				
RMSE													
Test Length*	+	+	+	+	904.73	135.72	15.22	86.03	1	.65	.22	.03	.15
Discrimination Level*	+	+	+	+	2.99	80.87	116.37	109.36	4	.02	.41	.49	.45
TL× DL	+	+	+	+	0.85	1.13	1.69	2.42	4	.01	.07	.07	.07
Error	†	†	†	†					490				

Note: RMSE: Root mean square error, TL: Test Length, DL: Discrimination Level,
MS: Mean square error,$+$: $MS < .0001$, † : $MS < .05$, df: degrees of freedom, *p<.001.

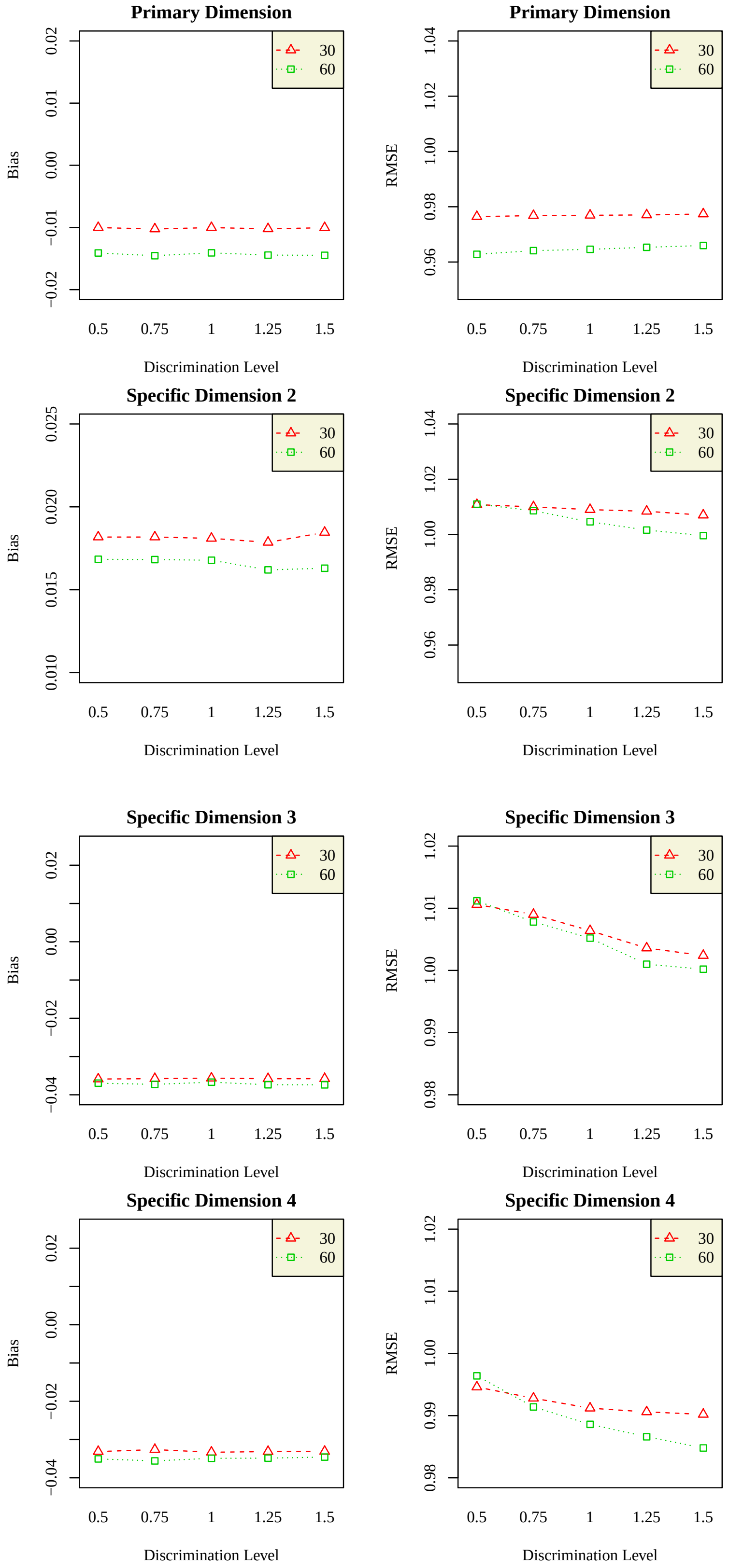


Figure 4.20: Interaction Plots Test Length and Discrimination Level on Bias and RMSE for Each Theta in BF-PC Model

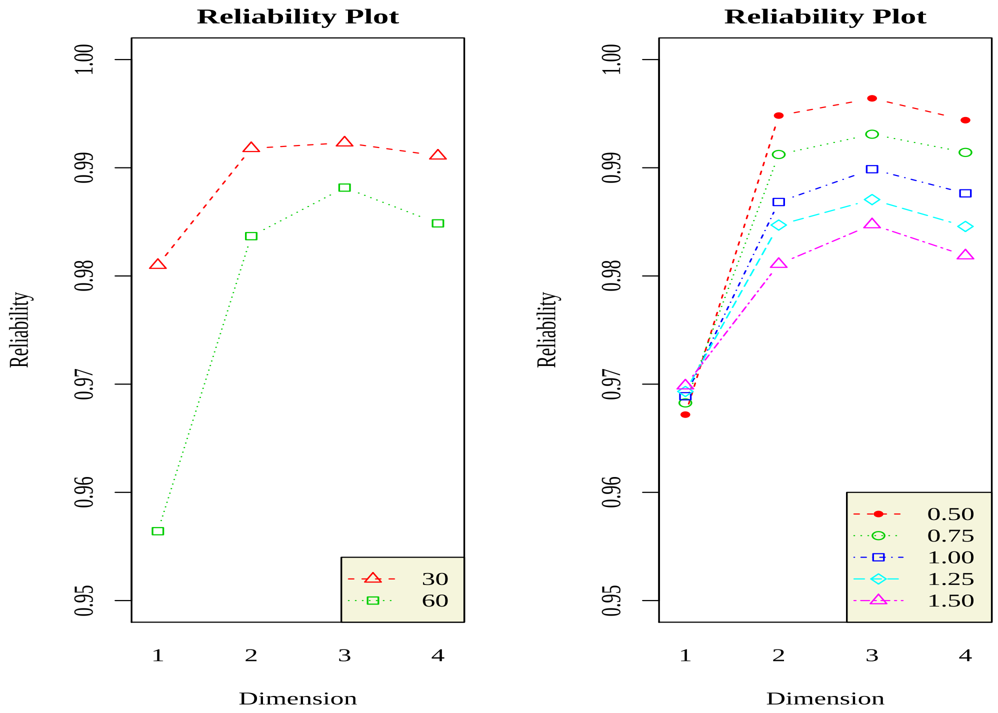


Figure 4.21: Reliability for Estimated Subscores and Generated Abilities (Thetas)

rameters over 50 replications. The simulation correlations in all cases for all ability dimensions were very high that is greater than .95 where the correlations for the specific ability dimensions were higher than the primary dimension. This study observed a higher simulation correlation in all ability dimension for the 30-item test compared to the 60-item test. As level of discrimination increases the correlation of the ability in the primary was not changed but correlations in the secondary ability dimensions decreased on a very small scale different.

4.3.3 Bayesian Criteria Comparison for Study 2

This section demonstrate comparisons of the Bayesian criteria for model complexity and fit at different test lengths and discrimination levels. The purpose of this section is solely to examine and demonstrate model-data fit across the two factors rather than for selecting or comparing the conditions in each or across factors.

Table 4.16 summarizes deviance, pD, DIC, AIC, BIC and -2 log Likelihood separated from different test lengths and discrimination levels. The results show that there were stable changes within each studied factor in all fit indices. Bayesian complexity and fit showed in Table 4.16 showed all the complexity of the 60-item test was as twice as big the 30-item test. This supports the MCMC estimations for a longer test to have more complex in terms of convergence of the MCMC that the estimation for all monitored parameters required more times than the shorter test.

Table 4.16: *BF-PC Model Bayesian Complexity and Fit*

Test Length		Discrimination Level				
		.50	.75	1.00	1.25	1.50
30	deviance	60412.16	60356.67	60313.90	60329.30	60336.07
	pD	1505.17	1548.03	1520.59	1500.70	1479.32
	DIC	61917.33	61904.70	61834.49	61830.00	61815.39
	AIC	63422.51	63452.74	63355.08	63330.70	63294.71
	BIC	65192.72	65273.35	65143.43	65095.66	65034.53
	-2 Log L	58906.99	58808.64	58793.31	58828.60	58856.75
60	deviance	120835.36	120782.71	120739.55	120706.73	120673.93
	pD	3058.66	3065.14	3092.13	3227.70	3228.84
	DIC	123894.03	123847.84	123831.68	123934.42	123902.77
	AIC	126952.69	126912.98	126923.80	127162.12	127131.61
	BIC	130549.9	130517.9	130560.4	130958.2	130929.00
	-2 Log L	117776.70	117717.60	11764.40	117479.00	117445.10

The distributions of the discrimination parameter that were varied in their discriminating levels between the primary and secondary ability dimensions did not much influence the MCMC convergence. Figure 4.22 can be referred to illustrate these results. The plots showed the differences in model complexity and fit from shorter to longer test across five levels of discrimination that smaller deviances were observed when the level of discrimination increases. There were fluctuations in this model complexity and fit when discrimination level of the secondary dimensions lower or higher than the primary dimension (i.e. at 1.00.)

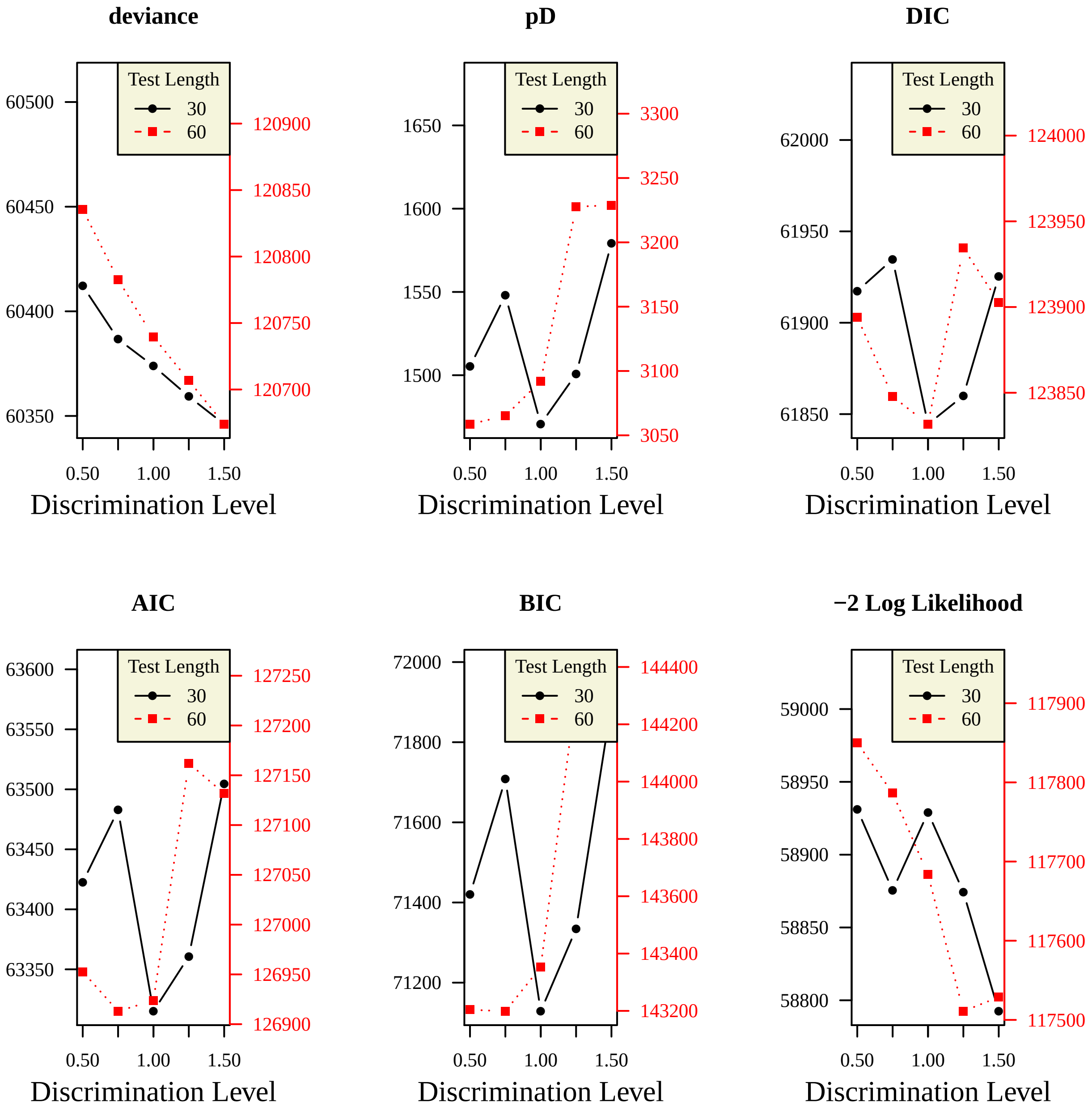


Figure 4.22: Bayesian Indices for BF-PC Model Complexity and Fit

4.3.4 Subscore Reliability and Classification

4.3.4.1 Subscore Reliability

Results from the accuracies of thetas estimates showed high reliabilities between the estimated subscores and the true subscores (i.e. reliabilities ranged from .95 to 1.00). In this section, Bayesian marginal reliabilities are computed from all estimated subscores. Subscore reliability based on the Bayesian marginal reliability for each condition in this model are summarized in Table 4.17.

For both test lengths, subscore reliabilities were observed to be considered highly reliable $(i.e.\hat{\rho} > .70)$ in all dimensions when the discrimination level of the specific ability dimension has equal or greater than the primary dimension. Subscores for the 30-item test in the specific ability dimensions were observed to have lower reliabilities than the longer 60-item test that ranged from. As expected, both shorter and longer test have the highest subscore reliabilities in the primary ability dimension, that is ranged from .90 to .98, as this dimension consists of all test items compared to specific ability dimensions that have shorter clusters of test items. When test is twice as long as the 30-item test, subscore reliabilities were improved as a function of discrimination level, that is ranged from .74 to .98.

Table 4.17: *Subscore Reliability from BF-PC Model*

Test Length	Discrimination Level				
Dimension	.50	.75	1.00	.125	1.50
30 Items					
1	.95	.93	.94	.98	.97
2	.60	.69	.73	.80	.89
3	.68	.63	.76	.81	.82
4	.66	.73	.76	.82	.86
60 Items					
1	.90	.90	.90	.90	.98
2	.74	.80	.81	.86	.95
3	.75	.82	.87	.87	.98
4	.78	.82	.86	.85	.95

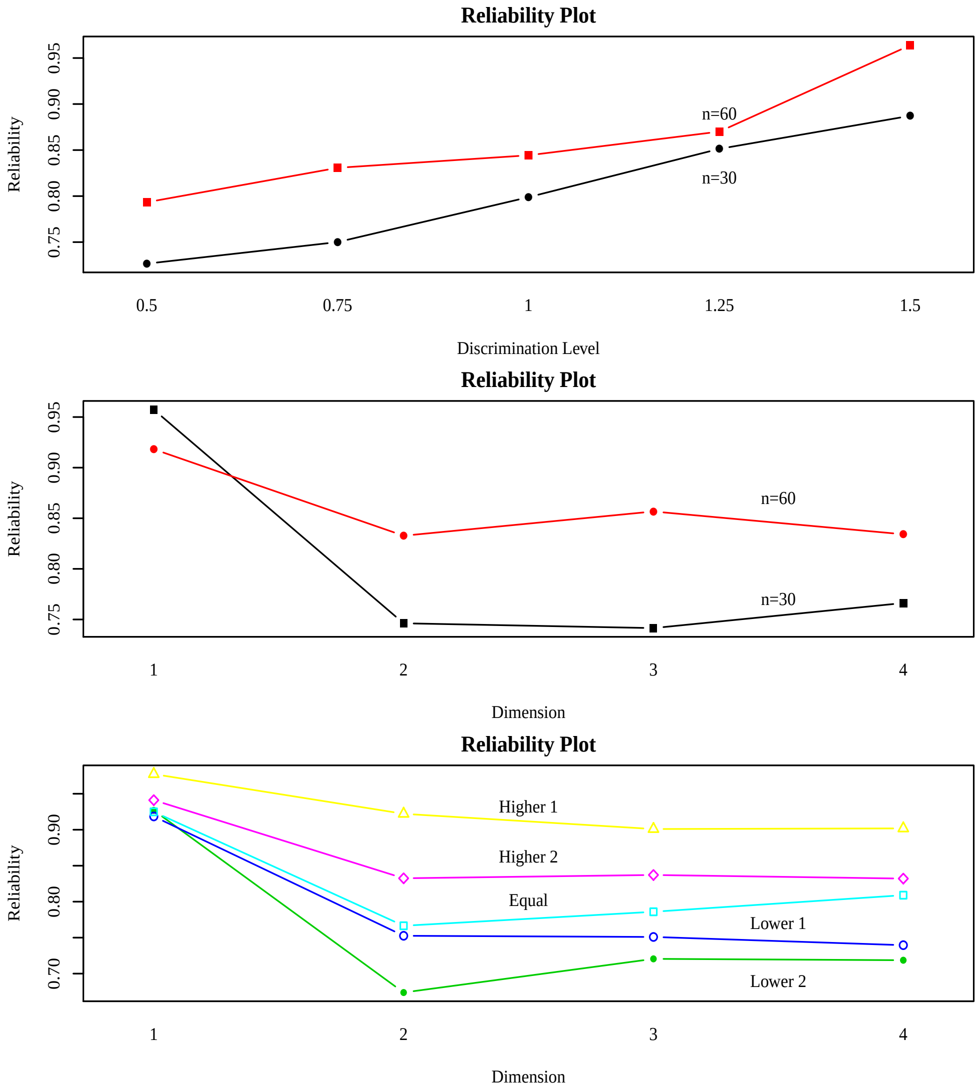


Figure 4.23: Subscore Reliability for BF-PC Model

In overall, a longer test with higher discrimination level in the specific ability dimension to the primary ability dimension showed higher subscore reliability than a shorter test. Figure 4.23 illustrates these results. The last plot clearly showed high subscore reliabilities that did not much change across specific ability dimensions but shown higher reliabilities when discrimination level increases.

4.3.4.2 Subscore Separation Index

To investigate further, subscore separation index (SSI) was computed to evaluate the quality of the subscores from the specific ability dimension that may be used for test reporting purposes. Table 4.18 summarizes the SSI that is also illustrated in Figure 4.24.

Test with 30 items showed lower SSI than tets with 60 items and SSI goes up when discrimination level increases. The SSI was ranged between .31 and 1.34 for 30-item test and between .41 and 2.51 for 60-item test. The associated hit rate for SSI is linearly related and presented as percentage of examinees have SSI greater than 1.0.

This model observed that, as discrimination level increases, not more than 1.5% of the examinees demonstrated SSI greater than 1.00 in the 30-item test the hit rate increased to 2.5% when the test is twice longer. Also, these hit rates showed that about more than 1% (up to less than 1.5%) of the examinees were showed SSI greater than 1.00 when the the level of discrimination of the specific ability dimension was equal or greater than the primary ability dimension.

Figure 4.24 clearly illustrates a higher SSI (and hit rate of SSI) for a longer test and at a higher level of discrimination. This result infers that the specific ability dimension subscores are separated from the primary ability dimension, that the SSI was 1 level greater than the primary ability dimension, when discrimination level increases or when a longer test is administered.

Table 4.18: *Subscore Separation Index Means, Standard Error and Hit Rate for BF-PC Model*

	Discrimination Level														
Test Length	.50			.75			1.00			1.25			1.50		
	1-2	1-3	1-4	1-2	1-3	1-4	1-2	1-3	1-4	1-2	1-3	1-4	1-2	1-3	1-4
SSI															
30 *Mean*	.004	.003	.004	.006	.005	.006	.009	.008	.010	.010	.010	.011	.013	.012	.013
s.e	.003	.002	.003	.004	.003	.004	.005	.004	.005	.005	.004	.005	.007	.005	.006
60 *Mean*	.007	.004	.007	.012	.008	.011	.017	.012	.015	.021	.016	.020	.025	.018	.023
s.e	.004	.003	.004	.005	.004	.005	.008	.006	.006	.008	.006	.006	.010	.007	.008
Hit Rate															
30 (%)	.37	.31	.44	.59	.53	.63	.91	.78	.95	.96	1.01	1.09	1.27	1.21	1.34
60 (%)	.66	.41	.68	1.16	.85	1.08	1.72	1.24	1.52	2.10	1.58	2.00	2.51	1.84	2.29

Note: SSI: Subscore Separation Index.

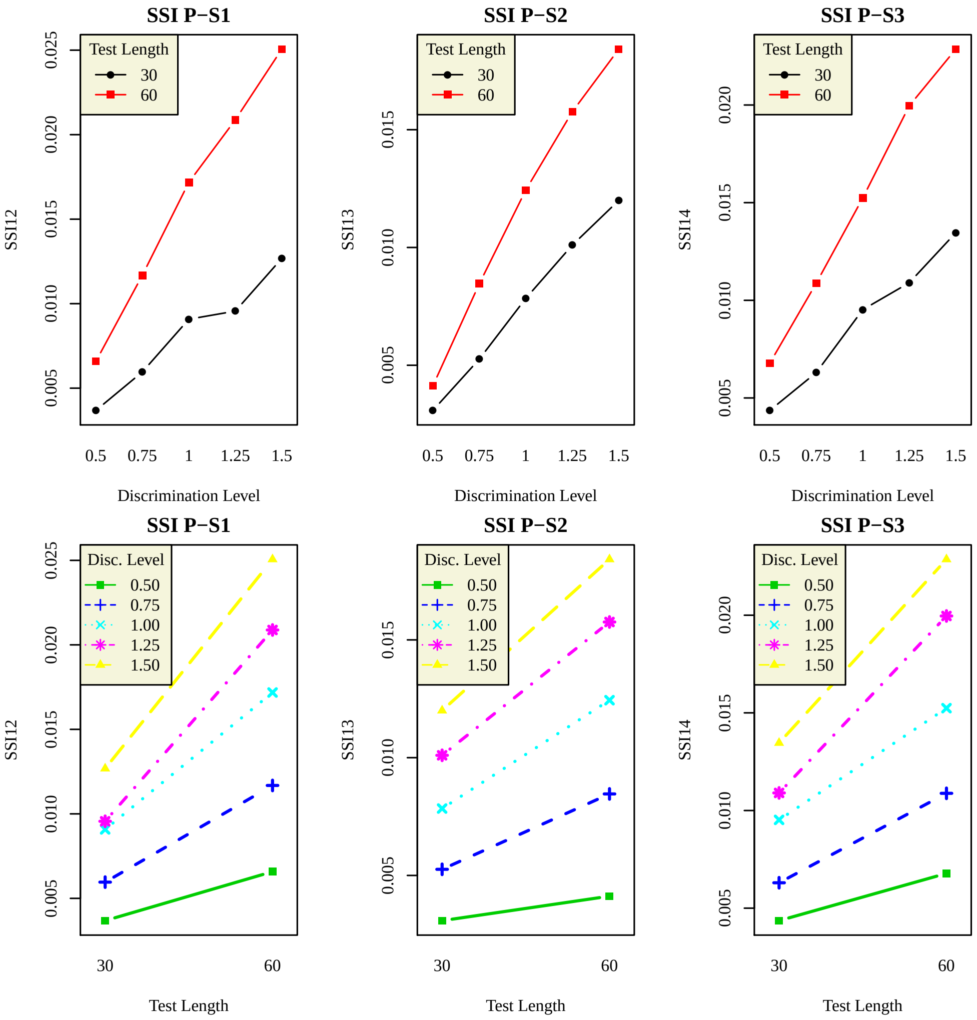


Figure 4.24: Plot of SSI and Hit Rate for BF-PC Model

Chapter 5

Discussion and Conclusion

The simulation studies in this dissertation were carried out to, primarily, estimate subscores reliability as well as to obtain classification of the subscores when subscores were estimated simultaneously from models assuming a test that is in nature come from a multidimensional structure. The bi-factor confirmatory multidimensional item response theory modeling is chosen for subscores estimation, reliability, and classification. Both compensatory and partially compensatory models are paired with the bi-factor model to consider the relationship between multiple dimensions of abilities being measured in a test, that an examinee's ability might be formed from a linear additive or nonlinear multiplicative function of multiple abilities. The effects of test length and discrimination level of test items in the specific ability dimension compared to the primary ability dimension were the focal designs in the simulation studies.

This chapter summarizes and discusses findings for the simulation studies. Each of the research question addressed in Chapter 1 is revisited and discussed based on the results in Chapter 4. Implications of this study to practical psychometric issues on improving subscores reliability and classification are discussed and concluded from the findings. Finally, limitation and direction for future research are addressed in the last section of this chapter.

5.1 Models for Improving Subscores Estimation

The first question stated in Chapter 1 of this dissertation was concerned about how the compensatory multidimensional item response theory and partially compensatory multidimensional item response theory models can be defined under the bi-factor model specification for subscores estimation, reliability, and classification. As defined in Chapter 3, the bi-factor compensatory multidimensional item response theory (BF-C) modeling is specified to simultaneously estimate subscores from multiple abilities in which one high ability can compensate any low abilities within a test. Thus, an examinee is highly likely to get an item correct when the examinee has a higher proficiency in one of the abilities being measured in test. Contrariwise, for the bi-factor partially compensatory model (BF-PC) that is from a nonlinear multiplicative modeling, being substantially high in one ability dimension can minimally compensate being low in other ability dimension. Therefore, an examinee is likely to get an item correct when the examinee has high proficiencies in all abilities being measured. As a matter of fact, the overall probability to answer an item correctly is bounded by the lower ability.

In addition to the nature of the compensation or partially compensation of abilities summarized above, all items in a test are best measured a primary ability and clusters of items are uniquely measured one or more specific (secondary) abilities. Test items are distinguished from each other when they are uniquely measured specific abilities or skills or objectives of the test. This specification can be defined from the bi-factor framework Gibbons et al. (1990); Gibbons & Hedeker (1992); Gibbons et al. (2007) from the degrees of discrimination of an item has to be separated from the primary ability dimension and item discriminations for other specific ability dimensions are constrained to zero. Since all items are best measured from the primary ability dimension, any shared variances among different abilities within a test are explained in this ability dimension, therefore, an overall subscore from this ability can be reported. Whereas, there are one or more secondary or specific ability dimensions that each of which subscore is reported to uniquely explain proficiency in the respective ability dimension. Primary and specific ability dimensions are all orthogonal to one another. Thus, an examinee will have two or more subscores from a test, one is for the primary (or overall) ability and other subscores are for specific content domains, skills or objectives.

Bayesian framework for the estimation of test item characteristics such as the discrimination, difficulty and guessing parameters together with abilities (thetas) are obtained from the Gibbs sampling of the Markov Chain Monte Carlo (MCMC) method. Despite computing difficulties and long waiting time in the MCMC simulation, the motivation of using this method is that rather than separate subscores estimations, the subscores for primary and specific ability dimensions are simultaneously estimated from the likelihood function of the studied models, that the subscores are also conditionally simultaneously estimated from the characteristics of the test items including the level of discrimination, difficulty level(s) as well as item guessing. Thousands of random draws for each parameter in the studied models (BF-C and BF-PC) from their respective candidates of the statistical distributions are considered in the Gibss sampling. The final convergent values are then considered to obtain the posterior distributions so as to compute the point estimates such as mean, mode or median and credible intervals for each parameter that are used to make statistical inference about the subscores such as reliability and classification.

In the next sections, summary and discussion of the simulation studies for the two studied models in this dissertation are presented.

5.2 Summary of Simulation Studies

5.2.1 Item and Person Parameters Recoveries

The first research question addressed in Chapter 1 was on how well the proposed models perform in recovering item and examinee parameters under various simulation conditions? Results from the simulation studies were averaged over the 50 replications of each simulation condition at successively convergent MCMC simulations.

Bias and RMSE for each monitored parameters were extensively examined separately for the BF-C and BF-PC models. Repeated measures ANOVA, a series of ANOVAs and ANOVA were carried out using SAS PROC GLM procedure to test significant different in bias and RMSE from the effects of test length and discrimination level. The simulation studies observed negative bias in the a-parameters for the BF-C model, whereas the recovery for the a-parameters in the BF-PC were positively biased. There were very small significant effects of test length for both models. The discrimination level was observed to largely effects bias and RMSE of the a-parameters. It was observed that as the discrimination level increases the bias and RMSE of a-parameters for both models decreases. The difficulty parameter for the BF-C and BF-PC models were not recovered well that bias and RMSE were quite large (greater than 1.00). Moreover, bias and RMSE of b-parameters of the BF-PC showed a better recovery with no significant effects of test length or discrimination level compared to the BF-C model that showed unstable recovery across discrimination levels with very small significant effects of test length. There was no significant effect of the five levels of discrimination on the recovery of b-parameters. The pseudo-guessing parameter in the BF-C model was very well recovered and there was considerably good recoveries of pseudo-guessing parameters in the BF-PC model that the bias and RMSE were closed to zero with no effect of test length or discrimination level.

These results showed that only pseudo-guessing parameters in BF-C and BF-PC models were recovered very well. Potentially, the result is related to the prior settings in the MCMC simulation for each model that the prior for c-parameter has been controlled to have a very informative prior. This is done so that the lower asymptote did not increase estimation complexities of the BF-C and BF-PC models. Although difficulty parameters were not of interests in both studies because the difficulty parameters in multidimensional item response theory framework typically left out as parameters to be estimated without constrained (Reckase, 2009), more studies on difficulty parameters in Bayesian framework could improve the recovery of this parameter. Finally, the simulation studies found substantial effects of the different levels of discrimination on the specific ability dimensions to the primary ability dimension. In all cases, better discrimination parameters recoveries were resulted as a function of discrimination level. The higher the discrimination level the better the recovery of the a-parameters. Moreover, the recoveries of the longer 20-item subtest showed very small to no different than the shorter 10-item subtest.

The abilities (thetas) parameter in both BF-C and BF-PC models were also very well recovered. The biases were very closed to zero with the associated error variances were closed to 1.0. There were unnoticeable effects of test length and discrimination level on the bias and RMSE of thetas. Moreover, Pearson's coefficients between the estimated and the true generated thetas were very high in the specific ability dimensions $(r > .90)$ in both models. There were lower thetas correlations in the primary ability dimension from the BF-C model, however, the correlations increases as a function of discrimination level. Thetas correlations from the BF-PC were very high that are closed to .98 or .99 for the all varying test lengths and discrimination levels.

5.2.2 Bayesian Model Complexity and Fit

This section is devoted to the second research question of this dissertation about model-data fits under the Bayesian framework. In overall, results of the Bayesian complexity and fit presented in Chapter 4 for BF-C model yield minimal decreasing in the deviance, effective number of parameters (pD), DIC, AIC, BIC and -2 log likelihood when the level of discrimination increases. Also, there were larger values in the Bayesian indices for the longer test compared to the shorter test. Thus, for the BF-C model, the complexity of the MCMC simulation increases as a function of test length and slightly influenced by the distribution of the discrimination levels of the specific ability dimensions to the primary ability. These observations explained the resulting waiting time addressed in the beginning of Chapter 4 that the 60-item test took one and a half day longer than the 30-item test for a converged MCMC simulation.

In the BF-PC model, there were fluctuations in model complexity and fit at varying discrimination levels and test lengths. Although the deviance were slightly decreases when the discrimination level increases, most of the changes in model complexity and fit were demonstrated whenever the discrimination level of the specific ability dimension equal to the discrimination level of the primary ability dimension (i.e. at 1.00). Apparently, the complexity and fit for the longer 60-item test was twice more than the 30-item test to support the observed longer time for the MCMC simulation in all studied conditions.

These reported results of Bayesian complexity and fit might not be used for model comparison or selection between the two studied models - BF-C and BF-PC. This is because the two models were designed and examined separately in the simulation studies. However, if the studies were intended for models comparison and/or selection, the presented results from the Bayesian complexity and fit perspectives can be used for, for example, when comparing one-, two-, and three-paramater bi-factor compensatory or bi-factor partially compensatory models as well as for nonnested models.

5.3 Subscores Reliability and Classification

Results from subsection 4.2.4 and subsection 4.3.4 are used to make inference about subscores reliability and classification, so as to answer the last research question of this dissertation. Results that are presented in Chapter 4 showed that there were substantially high Bayesian marginal subscores reliabilities, in average $\hat{\rho} > .90$ for the BF-C model and $\hat{\rho} > .80$ for the BF-PC model, respectively. For both models, higher subscores reliability resulted from lower bias and reduction in the error variance (i.e. RMSE) of thetas in all dimensions which is arrived at a higher discrimination level or for a longer test length.

Subscores reliabilities in the specific ability dimensions (dimension 2, 3, and 4) were somewhat lower than the subscore for the primary ability in the BF-PC model. Subscores reliabilities in the specific ability dimensions from the BF-C model showed equally high subscores reliability to the primary ability dimension. These results support BF-C as a better model that improved subscores reliabilities in the specific ability dimensions and maintain high reliability in the primary ability dimension.

The results also support a longer test to have higher subscores reliabilities. Also subscore reliabilities are improved when the levels of discrimination in the specific ability dimensions are higher than the primary ability dimension. This result supported the hypothesis that high discrimination level in the specific content domains improves the reliability of subscores.

The subscore separation index (SSI) that is greater than 1.00 was used as a benchmark to quantify the quality of the estimated subscores in the specific ability dimensions in terms of on how those subscores are highly separated from the primary ability dimension. Both studied models showed that SSI increases for a longer test and at a higher level of discrimination. Note that the SSI for the simulation studies only accounted for any values greater than 1.00 that the quality of subscores was determined based on subscores in the specific ability dimensions that were 1 level distinct (i.e. higher or lower) than the primary ability dimension. There were SSIs less than 1.0 but greater than 0.0 to distinguish subscores in the specific ability dimension from the primary dimension, which was not examined in this dissertation.

There were very low hit-rates of $SSI > 1.00$ from the BF-PC model, that ranged from .03 to 2.5 percent. This result may explain the nature of the partially compensatory model that the model did not substantially distinguish the subscores in the specific ability dimensions from the primary ability dimension. Thus, the subscores reported from BF-PC model make less distinction if they were estimated from the undimensional approach. This also implies that an examinee needs high performances in all abilities being measured in an item to answer the item correctly, which is also true that there were integrations of abilities that solely dominant by one primary ability within a person to perform well in a whole test. Another explanation from this results is that the choice of $SSI > 1.00$ for this model might be too strict due to the nature of the model that required all high proficiencies for better performance in a test. Thus, the index of separation between the specific ability dimension and primary ability dimension that is more conservative can be considered, for example $SSI > .30$ or $SSI > .50$.

SSIs for the BF-C model at varying simulation conditions were observed to be higher than the SSIs from the BF-PC model, that the hit-rates ranged from 7 to 15 percent and there were always SSIs greater than 0.00. Thus, the estimated subscores in the specific ability dimensions of the BF-C model showed 1 level distinct (i.e. higher or lower) than the subscore for the primary ability dimension. This implies that the BF-C model can be considered to explain many integrations of abilities that might be needed for an examinee to answer an item correctly, as well as to perform better in a test when the examinee has higher abilities in the specific ability dimensions than the primary ability dimension.

The hit-rates that are resulted from the SSI can be used to explain the frequencies of examinees with distinct subscores, that is higher or lower scores, in the specific ability dimensions compared to the primary dimension. Results from BF-PC model showed that there were less than 2 percent of the examinees had have dominant specific ability dimensions that had helped them to perform better in a test. From 1,500 examinees, this is equivalent to about 30 examinees that were recognized with their distinctions in the specific abilities that they have showed good or bad performance in a test with multiple dimension abilities. Whereas, 98 percent of the rest of the examinees may or may not have 1 level distinction (higher or lower) in their specific ability dimensions compared to the primary ability.

If BF-C model was considered for score reporting, there were more than 6 percent and up until 15 percent of the examinees had showed distinct performances in their specific ability dimensions than in the primary ability dimension. From 1,500 examinees, these percentages equivalent to about 90 and up to 210 examinees that had showed notable different in their performances that affected more from their specific ability dimensions than their primary ability, which implies their distinction performances in a test with multidimensional structure.

5.4 Limitations of the Study and Future Research

This study has several limitations. There were only two test lengths and five levels of discrimination considered for the estimations subscores, reliability and classification from the two defined models. The test lengths choose for these studies only showed small effects on the accuracies in the parameters estimation and both 10 and 20 items per subtest showed high subscores reliabilities. Different test lengths to represent shorter or longer subtests (e.g. 5, 15, 25 items) could be considered from the same models to investigate optimum number of items per subtest for reliable and valid subscores.

Also, this study was considered the specific ability dimensions to have discrimination levels that are two levels lower (.50 and .75), two levels higher (1.25 and 1.50) and an equal level of discrimination (1.00) to the primary ability dimension. Higher levels of discrimination demonstrated higher subscores reliability and classification. Different levels of discrimination could also be examined to differentiate the degrees of separation between the specific content dimensions and the primary dimension such as asking a question on how low (or high) the item discrimination would extremely effect the subscores reliability and classification when estimated from the two models?

In addition, this study examined only a sample size of 1,500 examinees and four dimensions of abilities. Smaller or larger sample sizes (e.g. 1,000, 1,200, 3,000 and 5,000) and more than four dimensions of ability could also be investigated to study the effects of sample size and number of dimensions on subscores reliabilities and classification. This could help to determine a tenable set of benchmark for the sample size and number of dimensions when estimating subscores from the BF-C and/or BF-PC models.

Different test lengths, discrimination levels, sample sizes, and number of dimensions can also be matched with an empirical example to better understand the BF-C and BF-PC models in estimating subscores, as well as to achieve reliable subscores for test reporting purposes or to quantify the classification of subscores that might be useful for diagnostics feedback on student's strengths and weaknesses, remedial action, educational placement or selection on college program and evaluation of the curriculum effectiveness. In term of Bayesian estimation from the BF-C and BF-PC models and if computing capacity or time became a concern, the estimation from empirical would not take longer than this two sets of 50-replication simulation study since there will be one MCMC simulation for each model and could be more depending on number of models or larger sample sizes, test lengths and dimensions to be considered. This direction for future research could be a good cross-validation study.

Another promising observation from this study if to further examine different levels of the index of separation between the specific ability dimension to the primary ability dimension. The *SSI* > 1.0 in this study might be too strict for one of the studied models. This study defined *SSI* in Chapter 3 as the standardized measure separation magnitude between a pair of subscores, and this is similar the Cohen's effect sizes analogy. Thus, *SSI* can be monitored at many different levels. Cut points of the *SSI* can be examined from low, medium to large level to reflect low, medium and large separation between the specific and primary ability proficiencies. Therefore, the highly reliable EAP subscores from BF-C and BF-PC models can be reported to examinees that the subscores represent their low, medium or high proficiency in the specific ability dimensions than the primary ability dimension that effects their performance in a test.

Kolen & Tong (2010) listed out several other psychometric properties that should be considered when estimating examinee's proficiencies from an item response theory framework. Two of which are the effects of estimator and effects of priors distributions on score distributions. For example, this study assuming the means of proficiencies for all dimensions were zeros and variances-covariances was an identity matrix. The means of proficiencies can be varied to examine the effects of different prior proficiency distributions on estimating the subscores, reliability and classification. The strengths of the priors for the item parameters could also be studied. Item parameters prior distributions from an empirical data could be emphasized for another good practical application. In general, for a simulation study, a very diffuse or non-informative set of priors could be used to reduce or control the effects of priors on the estimations of the EAP subscores. Guides on selecting priors in an IRT framework can be found in Beguin & Glas (2001), Fox & Glas (2001) and Gelman (2006).

Improving subscores reliability as well as subscores classification from the bi-factor and MIRT models can be wisely studied. Studies from the MIRT model has been widely appear in recent years due to its complexity for parameter estimations. Many studies are still looking for indeficiency answers from MIRT model that could better explain the multidimensional structures of mental abilities. Thus, the pairing of the bi-factor and MIRT models promises for more studies relating to subscores estimation, reliability and classification.

5.5 Practical Importance of the Study

In this simulation study, accuracy of parameter estimations in the specified bi-factor compensatory and partially compensatory models are better understood and can be referred to for future investigations about the efficacy of combining bi-factor and MIRT models for subscores evaluation as well as a contribution to new statistical methods for psychometric and educational measurement in test scoring and reporting.

The specification of bi-factor multidimensional item response theory from both compensatory and partially compensatory models using Bayesian approach, for apparently the first time, for subscores estimation, evaluating subscore reliability, and quantifying subscores classification is practically important for studies related to improving subscores reliability in large-scale assessment. The observed results under different simulation conditions help to understand the characteristics of a test, to evaluate or achieve reliable subscores by using the subscores reliability index from Bayesian framework, and to quantified the degrees of separation between primary and specific content domains using an index called *SSI*. Bayesian complexity and fit indices described in this study are important about model fitting of bi-factor compensatory and partially compensatory models in choosing a preferred model that better supports the characteristics of the designed test to the administered examinees.

The reliable and quantified subscores are from the well-defined bi-factor MIRT models. The reliable subscores and information provided in *SSI* might be useful for diagnostic purposes in terms of giving feedback to each examinee about their academic performance, strengths and weaknesses as well as for remedial actions or for future decision such as improving college or career readiness for the examinee. Reliable and quantified subscores for any subject areas such as mathematics, sciences and English Language Arts (ELA) are also important to states and institutions to prole their students' performance, to better evaluate their curriculum's effectiveness and to focus in areas that need remediation.

References

Ackerman, T., Gierl, M., & Walker, C. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22(3), 37–51.

Ackerman, T. A. & Davey, T. C. (1991). Concurrent adaptive measurement of multiple abilities. In *Proceeding of the National Council on Measurement in Education Annual Meeting* Chicago, IL.

AERA, NCME, & APA (1999). Standards for educational and psychological testing.

Babcock, B. (2011). Estimating a noncompensatory IRT model using metropolis within gibbs sampling. *Applied Psychological Measurement*, 35, 317–329.

Baker, Frank, B. (1985). *The Basics of Item Response Theory*. University of Marryland, College Park, MD: Heinemann and ERIC Clearinghouse on Assessment and Evaluaiton, second, revised and updated (2001) edition.

Baker, Frank, B. & Kim, S. (2004). *Item response theory: parameter estimation techniques.* New York: Marcel Dekker, Inc, second edition, revised and expanded edition.

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

Birnbaum, A. (1968). *Statistical theories of mental test scores*, chapter Some latent trait models and their use in inferring an examinee's ability., (pp. 397–472). Addison-Wesley.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-Information item factor analysis. *Applied Psychological Measurement*, 12(3), 261–280.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, 34(3), 197–211.

Bolt, D. M. & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using markov chain monte carlo. *Applied Psychological Measurement*, 27, 395–414.

Bradlow, E., Wainer, H., & Wang, X. (1999). A bayesian random effects model for testlets. *Psychometrika*, 64(2), 153–168.

Brooks, S., Gelman, A., Jones, G. L., & Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: Chapman & Hall/CRC.

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

Carlin, B. P. & Louis, T. A. (2000). *Bayes and Empirical Bayes methods for Data Analysis*. London: Chapman & Hall.

Casella, G. & George, E. (1992). Explaining the gibbs sampler. *American Statistician*, (pp. 167–174).

Chambers, J. M. (2008). *Software for Data Analysis: Programming with R*. New York: Springer.

Chib, S. & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, (46), 167–174.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Wadsworth Publishing.

Croudace, T., Ploubidis, G., & Abbott, R. (2005). Bilog-mg, multilog, parscale and testfact4 programs on cd with single volume manual combining reference guide and users guide (program examples: Input and output). *British Journal of Mathematical and Statistical Psychology*, 58, 193–195.

de la Torre, J. & Hong, Y. (2010). Parameter estimation with small sample size a Higher-Order IRT model approach. *Applied Psychological Measurement*, 34, 267–285.

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

de la Torre, J. & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A Higher-Order IRT model approach. *Applied Psychological Measurement*, 33(8), 620–639.

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement*, 35(4), 296–316.

DeMars, C. (2005). Scoring subscales using multidimensional item response theory models. In *Poster presented at the Annual Meeting of the American Psychological Association (Washington, DC, 2006)* (pp.$\tilde{2}$5). Washington, DC.

DeMars, C. E. (2006). Application of the Bi-Factor multidimensional item response theory model to Testlet-Based tests. *Journal of Educational Measurement*, 43(2), 145–168.

Deng, N., Wells, C., & Hambleton, R. (2008). : Connecticut.

Drasgow, F. & Olson-Buchanan, J. B. (1999). *Innovations in computerized assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Edwards, M. (2002). Augmenting IRT scale scores:a monte carlo study to evaluate an empirical bayes approach to subscore augmentation.

Edwards, M. C. (2010). A markov chain monte carlo approach to confirmatory item factor analysis. *Psychometrika*, 75(3), 474–497.

Edwards, M. C. & Vevea, J. L. (2006). An empirical bayes approach to subscore augmentation: How much strength can we borrow? *JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS*, 31(3), 241–259.

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

Gamerman, D. (1997). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. London: Chapman & Hall.

Gelman, A. (1996). *Markov Chain Monte Carlo*, chapter Inference and monitoring convergence. Chapman and Hall/CRC: London, eds edition.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian analysis*, 1(3), 515–533.

Gelman, A., Carlin, John, B., Stern, Hal, S., & Rubin, D. B. (2004). *Bayesian data analysis*. Boca Raton: Chapman and Hall/CRC.

Gelman, A. & Shirley, K. (2011). Inference from simulations and monitoring convergence andrew gelman. In *Handbook of Markov Chain Monte Carlo* (pp. 163–174). Boca Raton, FL: Chapman & Hall/CRC.

Geman, S. & Geman, D. (1984). Stochastic relaxation, gibbs distribution and bayesian restoration of image. *IEE Transaction on Pattern Analysis and Machine Intelligence*, (6), 721–741.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-Information item bifactor analysis of graded response data. *Applied Psychological Measurement*, 31(1), 4–19.

Gibbons, R. D. & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57(3), 423–436.

Gibbons, R. D., Hedeker, D. r., & Bock, R. D. (1990). *Full-Information Item Bi-Factor Analysis: ONR Technical report*. Technical report, Biometric Laboratory, Illinois State Psychiatric Institute, Chicago.

Gibbons, R. D., Rush, A. J., & Immekus, J. C. (2009). On the psychometric validity of the domains of the PDSQ: an illustration of the bi-factor item response theory model. *Journal of psychiatric research*, 43(4), 401–410.

Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). Introducing markov chain monte carlo. In *Markov chain Monte Carlo in practice* (pp. 1–17). London: Chapman & Hall.

Green, B., Bock, R., Humphreys, L., Linn, R., & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21(4), 347–360.

Haberman, Shelby, J. & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, (pp. 1–19).

Haberman, S. (2008). When can subscores have value? *JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS*, 33(2), 204–229.

Haberman, S., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, 62(1), 79–95.

Hambleton, Ronald, K. & Jones, Russell, W. (1993). Comparison of classical test theory an dItem response theory and their application to test development. *Educational Measurement: Issues and Practices*, 12, 38–47.

Hambleton, Ronald, K. & Swaminathan, H. (1985). *Item Response Theory*. Boston/Dordrecht/Lancaster: Kluwer.Nijhoff Publishing.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1), 97–109.

Holzinger, K. J. & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2(1), 41–54.

Immekus, J. C. & Imbrie, P. K. (2008). Dimensionality assessment using the Full-Information item bifactor analysis for graded response data: An illustrative with the state metacognitive inventory. *Educational and Psychological Measurement*, 68(4), 695–709.

Jöreskog, K. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2), 183–202.

Kahraman, N. & Thompson, T. (2011). Relating unidimensional IRT parameters to a multidimensional response space: A review of two alternative projection IRT models for scoring subscales. *Journal of Educational Measurement*, 48(2), 146–164.

Kelly, D. & Curtis, S. (2011). *Bayesian Inference for Probabilistic Risk Assessment: A Practitioner's Guidebook*, chapter More Complex Models for Random Durations, (pp. 89–109). Springer-Verlag: London.

Kelly, T. L. (1927). *Interpretation of Educational Measurements*. New York, NY: World Book Company.

Kelly, T. L. (1947). *Fundamentals of Statistics*. Cambridge: Harvard University Press.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer, 2nd edition.

Kolen, M. J. & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice*, 29(3), 8–14.

L'Ecuyer, P., Simard, R., Chen, E. J., & Kelton, W. D. (2002). An object-oriented random-number package with many long streams and substreams. *Operations Research*, 50(6), 1073–1075.

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement*, 25(4), 357–372.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, 30(1), 3–21.

Li, Y. & Rupp, A. A. (2011). Performance of the s2 statistic for Full-Information bifactor models. *Educational and Psychological Measurement*, 71(6), 986–1005.

Lord, F. (1952). A theory of test scores. *Psychometric*.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for dichotomous items. In *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons responses and performances as scientific inquiry into score meaning. *American Psychologist*, (pp. 741–749).

Metropolis, N. (1987). The beginning of the monte carlo method. *Los Alamos Science*, (Speacial Issue), 125–131.

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2002). *Computer-Based Testing: Building the Foundation for Future Assessments*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mislevy, R. (1987). Exploiting auxiliary infornlation about examinees in the estimation of item parameters. *Applied Psychological Measurement*, 11(1), 81.

Muthén, B. (1989). Factor strcuture in groups selected observed scores. *British Journal of Mathematical and Statistical Psychology*, (42), 81–90.

NCLB (2001). No child left behind act (nclb) of 2001.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2001). *Practical considerations in computer testing*. New York: Springer-Verlag.

Patz, Richard, J. & Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS*, 24(4), 342–366.

Patz, R. & Junker, B. (1999b). A straightforward approach to markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24(2), 146.

Pommerich, M., Nicewander, W. A., & Hanson, B. (1999). Estimating average domain scores. *Journal of Educational Measurement*, 36(3), 199–216.

Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2010). The utility of augmented sucbscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, 23, 266–285.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21(1), 25–36.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York: Springer.

Reise, S., Moore, T., & Haviland, M. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, 92(6), 544–559.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16(S1), 19–31.

Reise, S. P., Ventura, J., Keefe, R. S., Baade, L. E., Gold, J. M., Green, M. F., Kern, R. S., Mesholam-Gately, R., Nuechterlein, K. H., Seidman, L. J., et al. (2011). Bifactor and item response theory analyses of interviewer report scales of cognitive impairment in schizophrenia. *Psychological assessment*, 23(1), 245.

Rijmen, F. (2009). *Three Multidimensional Models for Testlet-Based Tests: Formal Relations and an Empirical Comparison*. Technical Report ETS RR-09-37, Educational Testing Service, Princeton, New Jersey.

Rijmen, F. (2010). Formal relations and an empirical comparison among the Bi-Factor, the testlet, and a Second-Order multidimensional IRT model. *Journal of Educational Measurement*, 47(3), 361–372.

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika*, 73(2), 167–182.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores.

Schmid, J. & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22(1), 53–61.

Schmid Jr, J. (1957). The comparability of the Bi-Factor and Second-Order factor patterns. *The Journal of Experimental Educational*, (pp. 249–253).

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

Shin, D. (2007). A comparison of method of estimating subscale scores for Mixed-Format tests. *Pearson Educational Measurement*.

Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, 26(4), 21–28.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of Testlet-Based tests. *Journal of Educational Measurement*, 28(3), 237–247.

Skorupski, W. P. & Carvajal, J. (2009). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement*, 70(3), 357–375.

Spiegelhalter, D. J., Best, N. G., & Carlin, B. P. (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models*. Research Report 98-009, Division of Biostatistics, University of Minnesota.

Spiegelhalter, David, J., Best, Nicola, G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistics*, 64(4), 583–639.

Stone, C., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Taylor & Francis Group*, (23), 63–86.

Stout, W., Ackerman, T., Bolt, D., Froelich, A. G., & Heck, D. (2003). *On the Use of Collateral Item Response Information to Improve Pretest Item Calibration*. Computerized Testing Report 98-13, Law School Admission Council.

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2openbugs: A package for running openbugs from r. *Journal of Statistical Software*, 12, 1–16.

Swineford, F. (1941). Some comparisons of the multiple-factor and the bi-factor methods of analysis. *Psychometrika*, 6(6), 375–382.

Tao, S. (2009). Using collateral information in the estimation of sub-scores—a fully bayesian approach. *Theses and Dissertations*, (pp. 321).

Tate, R. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education*, 17(2), 89–112.

Thissen, D. (1982). Maximum likelihood. *Psychometrika*, 47(2), 175–186.

Thissen, D. & Edwards, M. C. (2005). Diagnostic scores augmented using multidimensional item response theory: Preliminary investigation of MCMC strategies. In *annual meeting of the National Council on Educational Measurement, Montreal, Canada*.

Thissen, D. & Orlando, M. (2001). Item response theory for items scored in two categories. In *Test Scoring*.

Thompson, N. A. (2006). Item response theory parameterization of the multistate bar exam nathan a. thompson and david j. weiss university of minnesota.

Thurstone, L. (1947). *Multiple-factor Analysis*. Technical report, University of Chicago, Chicago, IL.

van der Linden, W. J. & Glas, C. A. (2000). *Computerized adaptive testing: Theory and Practice*. Dordrecht, The Nederlands: Kluwer Academic Publisher.

Wainer, H. & Dorans, N. J. (2000). *Computerized adaptive testing: a primer*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making praxis scores more useful. *Journal of Educational Measurement*, 37(2), 113–140.

Wainer, H., Vevea, J. L., Camacho, F., Reeve III, B. B., Rosa, K., Nelson, L., Swygert, K., & Thissen, D. (2001). Augmented scores - "Borrowing Strenght" to compute score based on small numbers of items. In *Test Scoring* (pp. 343–387). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wang, P. & Gao, F. (2010). Full-Information item bifactor analysis of the job burnout scale for chinese college teachers. 5(7), 155.

Wang, W., Chen, P., & Cheng, Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, 9(1), 116–136.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement*, 47(3), 339–360.

Yen, Wendy, M. (1984). Obtaining maximum likelihood trait estimates from Number-Correct scores for the Three-Parameter logistic model. *Journal of Educational Measurement*, 21(2), 93–111.

Yen, W. (1987). A Bayesian/IRT index of objective performance. In *annual meeting of the Psychometric Society, Montreal, Quebec, Canada, June* (pp. 1–19).

Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, 64(2), 113–128.

Appendix A

Appendices for Study 1

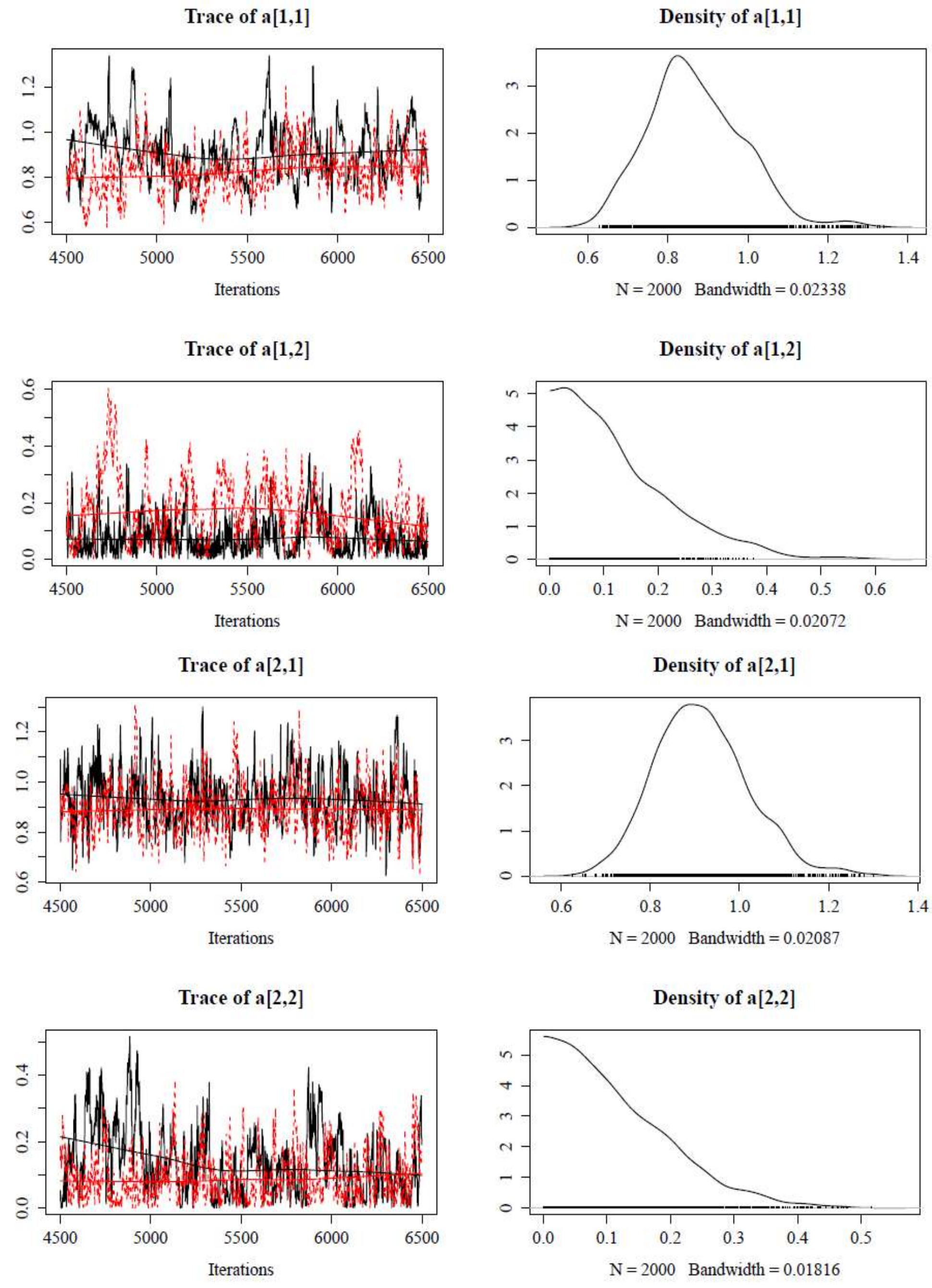
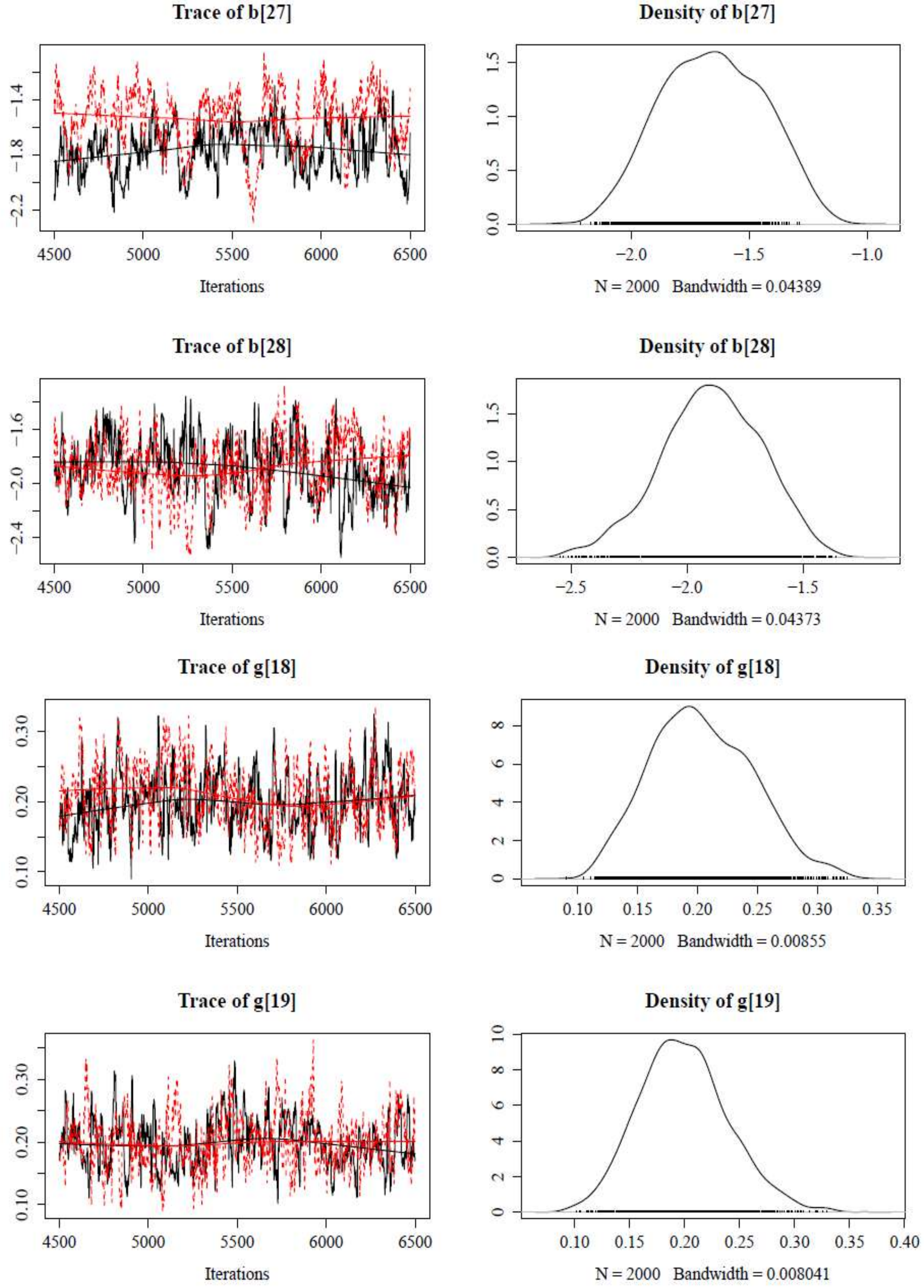


Figure A.1: Example Trace and Kernel Density Plots for a-parameter



Trace of b[27]

Density of b[27]

Trace of b[28]

Density of b[28]

Trace of g[18]

Density of g[18]

Trace of g[19]

Density of g[19]

Figure A.2: Example Trace and Kernel Density Plots for b- and c-parameter

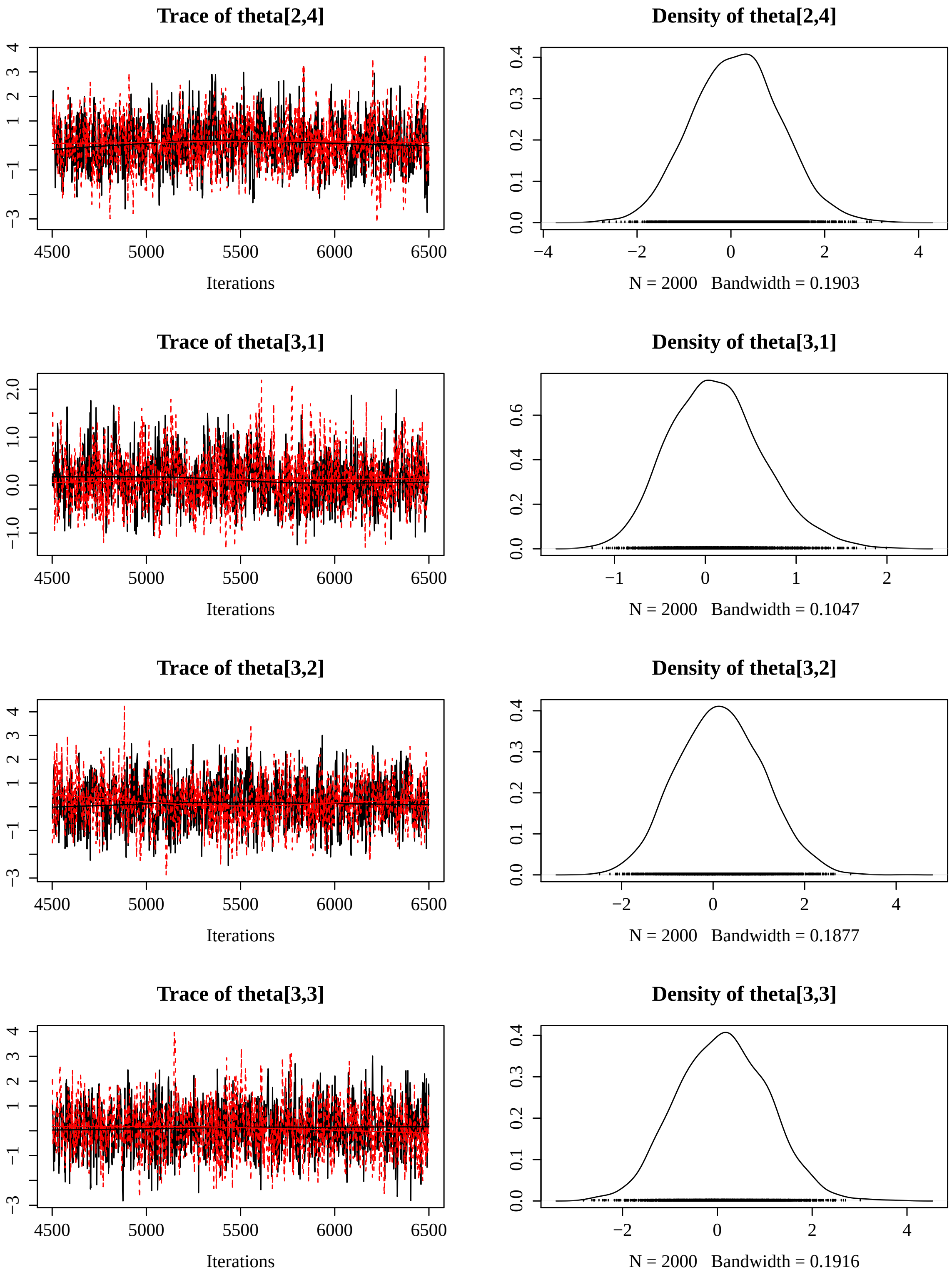


Figure A.3: Example Trace and Kernel Density Plots for Thetas

Table A.1: *Means and Standard Errors for Bias and RMSE of a-parameter*

			Bias		RMSE	
Test Length	Discrimination Level	d, dimension	*M*	*s.e.*	*M*	*s.e.*
30	0.50	1	0.156	0.218	0.352	0.108
		2	-0.048	0.105	0.757	0.108
		3	-0.056	0.100	0.760	0.093
		4	-0.047	0.110	0.753	0.117
	0.75	1	0.047	0.236	0.333	0.148
		2	-0.120	0.110	0.670	0.067
		3	-0.124	0.115	0.662	0.067
		4	-0.120	0.123	0.665	0.079
	1.00	1	-0.069	0.239	0.333	0.173
		2	-0.197	0.116	0.561	0.048
		3	-0.199	0.121	0.558	0.059
		4	-0.192	0.130	0.552	0.045
	1.25	1	-0.185	0.221	0.372	0.195
		2	-0.274	0.119	0.447	0.085
		3	-0.277	0.116	0.455	0.101
		4	-0.275	0.130	0.447	0.092
	1.50	1	-0.280	0.216	0.416	0.212
		2	-0.360	0.111	0.337	0.156
		3	-0.358	0.119	0.344	0.155
		4	-0.357	0.126	0.330	0.131
60	0.50	1	0.273	0.138	0.355	0.075
		2	-0.108	0.044	0.769	0.037
		3	-0.104	0.047	0.777	0.037
		4	-0.107	0.046	0.775	0.038
	0.75	1	0.187	0.137	0.300	0.083
		2	-0.187	0.054	0.644	0.033
		3	-0.185	0.051	0.652	0.030
		4	-0.185	0.061	0.647	0.038
	1.00	1	0.067	0.141	0.260	0.108
		2	-0.265	0.058	0.517	0.036
		3	-0.262	0.060	0.522	0.034
		4	-0.265	0.062	0.517	0.036
	1.25	1	-0.035	0.146	0.256	0.128
		2	-0.344	0.061	0.386	0.056
		3	-0.341	0.061	0.391	0.060
		4	-0.343	0.067	0.386	0.060
	1.50	1	-0.146	0.134	0.282	0.135
		2	-0.429	0.058	0.244	0.073
		3	-0.425	0.060	0.252	0.077
		4	-0.428	0.065	0.247	0.079

Table A.2: *Means and Standard Errors for Bias and RMSE of b-parameter*

Test Length	Discrimination Level	Bias		RMSE	
		M	*s.e.*	*M*	*s.e.*
30	0.50	0.686	0.739	2.211	1.690
	0.75	0.807	0.721	2.393	1.817
	1.00	0.842	0.752	2.503	1.937
	1.25	0.825	0.828	2.607	1.922
	1.50	0.755	0.877	2.562	1.934
60	0.50	0.527	0.612	1.649	0.979
	0.75	0.512	0.709	1.745	1.087
	1.00	0.706	0.571	1.818	1.119
	1.25	0.536	0.839	1.982	1.305
	1.50	0.593	0.646	1.774	1.163

Table A.3: *Means and Standard Errors for Bias and RMSE of c-parameter*

Test Length	Discrimination	Bias		RMSE	
		M	*s.e.*	*M*	*s.e.*
30	0.50	0.010	0.003	0.000	0.000
	0.75	0.009	0.003	0.000	0.000
	1.00	0.008	0.003	0.000	0.000
	1.25	0.008	0.006	0.001	0.001
	1.50	0.007	0.007	0.001	0.001
60	0.50	0.007	0.003	0.000	0.000
	0.75	0.007	0.003	0.000	0.000
	1.00	0.007	0.003	0.000	0.000
	1.25	0.007	0.006	0.001	0.002
	1.50	0.004	0.002	0.000	0.000

Table A.4: *Means and Standard Errors for Bias, RMSE, and Pearson's Correlation for Thetas*

TL	DL	*d*	Bias		RMSE		Correlation	
			M	*s.e.*	*M*	*s.e.*	*M*	*s.e.*
30	0.50	1	-0.006	0.005	0.682	0.126	0.670	0.147
		2	0.017	0.001	1.023	0.012	0.876	0.008
		3	0.011	0.001	1.015	0.016	0.894	0.008
		4	0.040	0.001	1.027	0.014	0.890	0.008
	0.75	1	-0.007	0.005	0.783	0.108	0.691	0.163
		2	0.017	0.001	1.019	0.014	0.917	0.009
		3	0.011	0.001	1.014	0.011	0.934	0.009
		4	0.039	0.001	1.013	0.010	0.930	0.009
	1.00	1	-0.007	0.005	0.864	0.084	0.730	0.135
		2	0.017	0.001	1.019	0.016	0.965	0.009
		3	0.011	0.001	1.012	0.013	0.982	0.008
		4	0.040	0.001	1.011	0.011	0.978	0.009
	1.25	1	-0.005	0.010	0.938	0.056	0.776	0.102
		2	0.017	0.002	1.017	0.014	0.960	0.009
		3	0.011	0.001	1.011	0.013	0.977	0.008
		4	0.040	0.001	1.004	0.008	0.973	0.009
	1.50	1	-0.004	0.015	0.980	0.051	0.820	0.090
		2	0.017	0.002	1.015	0.013	0.961	0.008
		3	0.011	0.001	1.009	0.013	0.978	0.008
		4	0.040	0.001	1.004	0.006	0.974	0.008
60	0.50	1	-0.002	0.010	0.602	0.100	0.713	0.112
		2	0.019	0.003	1.043	0.019	0.875	0.007
		3	-0.011	0.023	1.039	0.019	0.893	0.007
		4	0.035	0.005	1.019	0.017	0.889	0.007
	0.75	1	-0.001	0.012	0.726	0.091	0.677	0.172
		2	0.018	0.003	1.044	0.023	0.920	0.009
		3	-0.009	0.023	1.036	0.017	0.938	0.009
		4	0.036	0.005	1.014	0.017	0.934	0.009
	1.00	1	-0.004	0.008	0.834	0.068	0.688	0.133
		2	0.019	0.003	1.039	0.018	0.966	0.007
		3	-0.008	0.023	1.035	0.019	0.984	0.007
		4	0.036	0.004	1.007	0.012	0.980	0.007
	1.25	1	0.003	0.019	0.912	0.061	0.713	0.085
		2	0.019	0.003	1.037	0.018	0.964	0.008
		3	-0.009	0.022	1.033	0.017	0.982	0.008
		4	0.036	0.006	1.004	0.013	0.978	0.008
	1.50	1	0.002	0.014	0.966	0.053	0.767	0.074
		2	0.018	0.002	1.037	0.019	0.965	0.007
		3	-0.008	0.023	1.030	0.017	0.982	0.007
		4	0.036	0.005	1.004	0.013	0.978	0.007

Note: TL: Test Length, DL: Discrimination Level, *d*: dimension

Appendix B

Appendices for Study 2

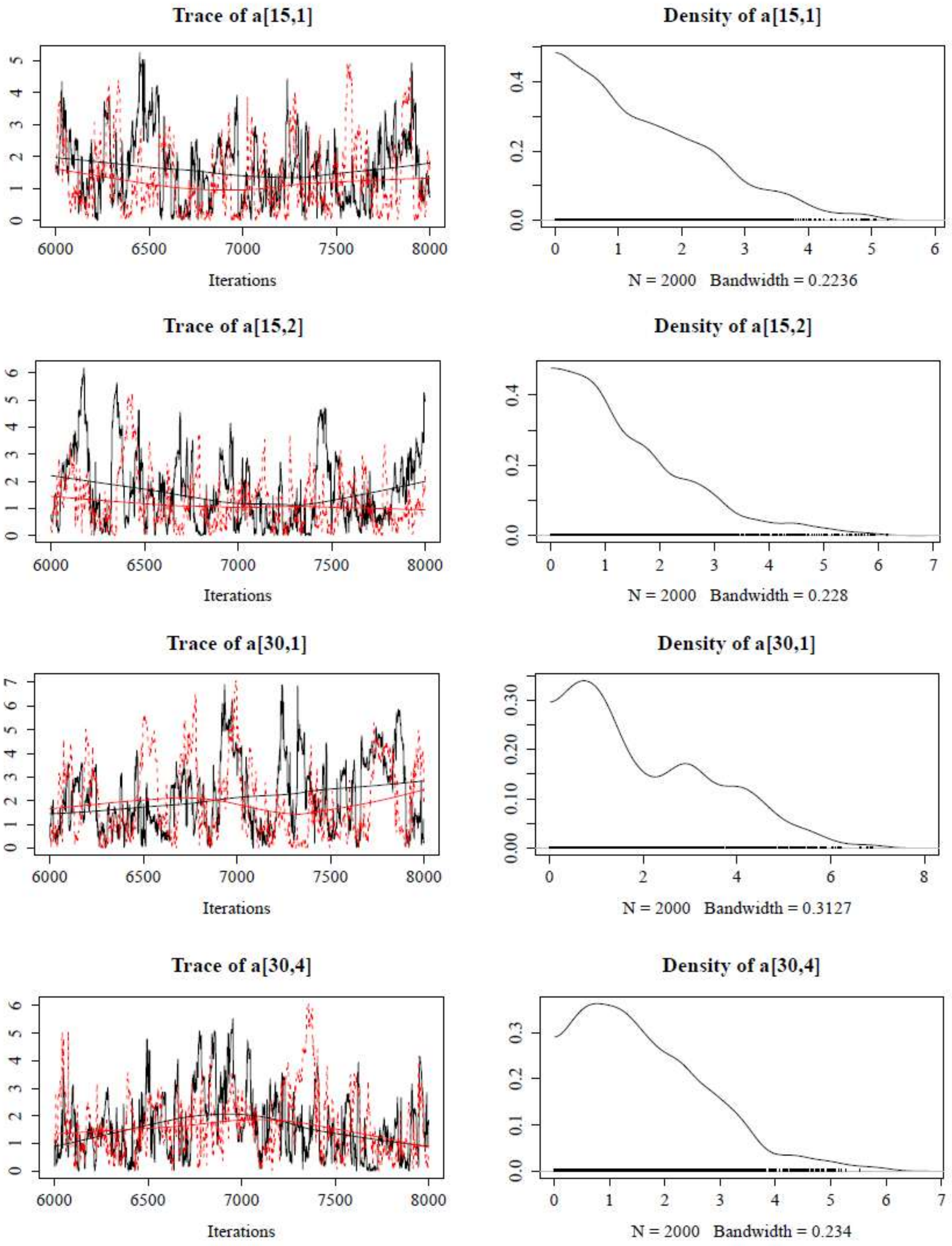


Figure B.1: Example Trace and Kernel Density Plots for a-parameter

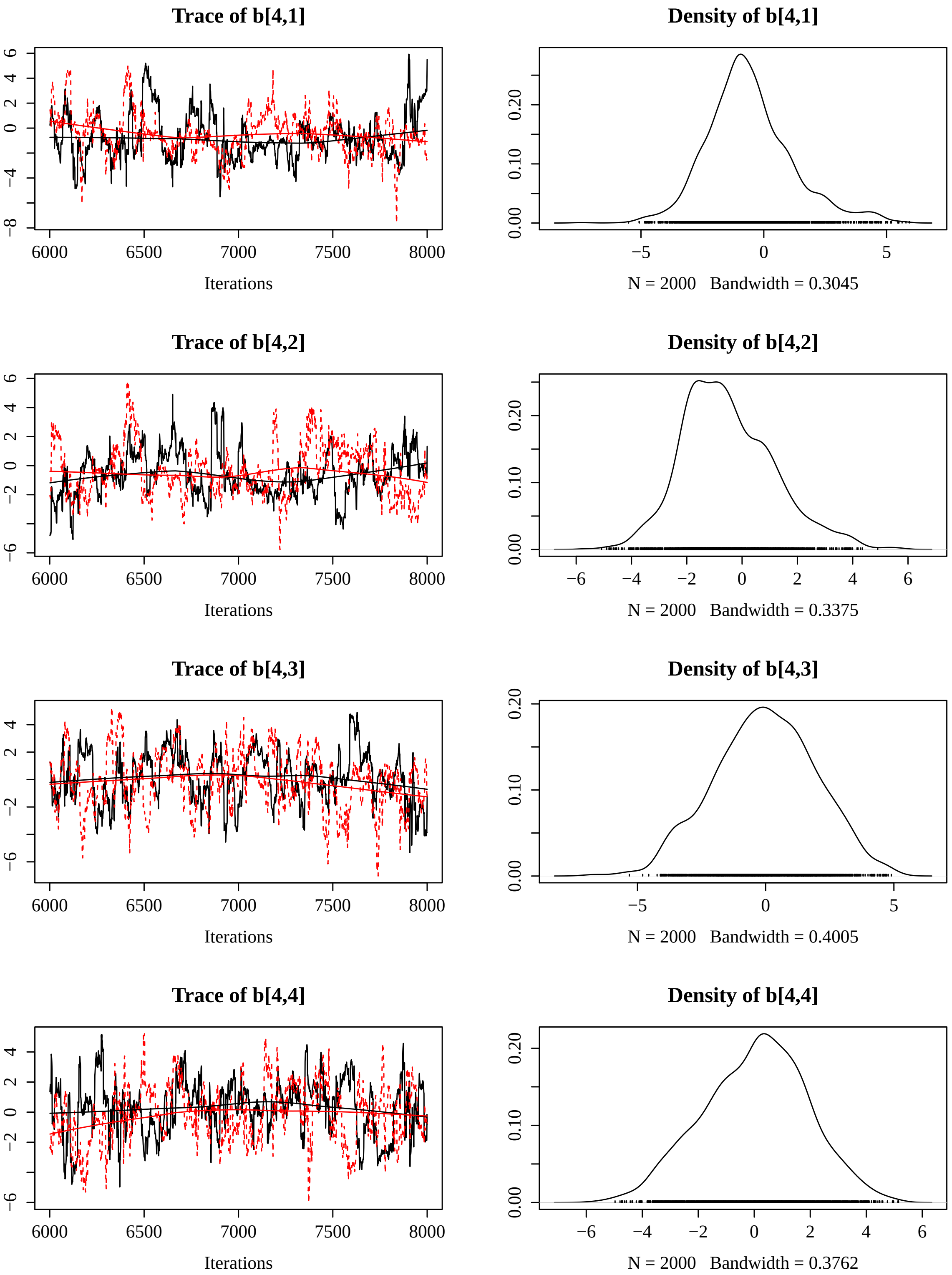


Figure B.2: Example Trace and Kernel Density Plots for b-parameters

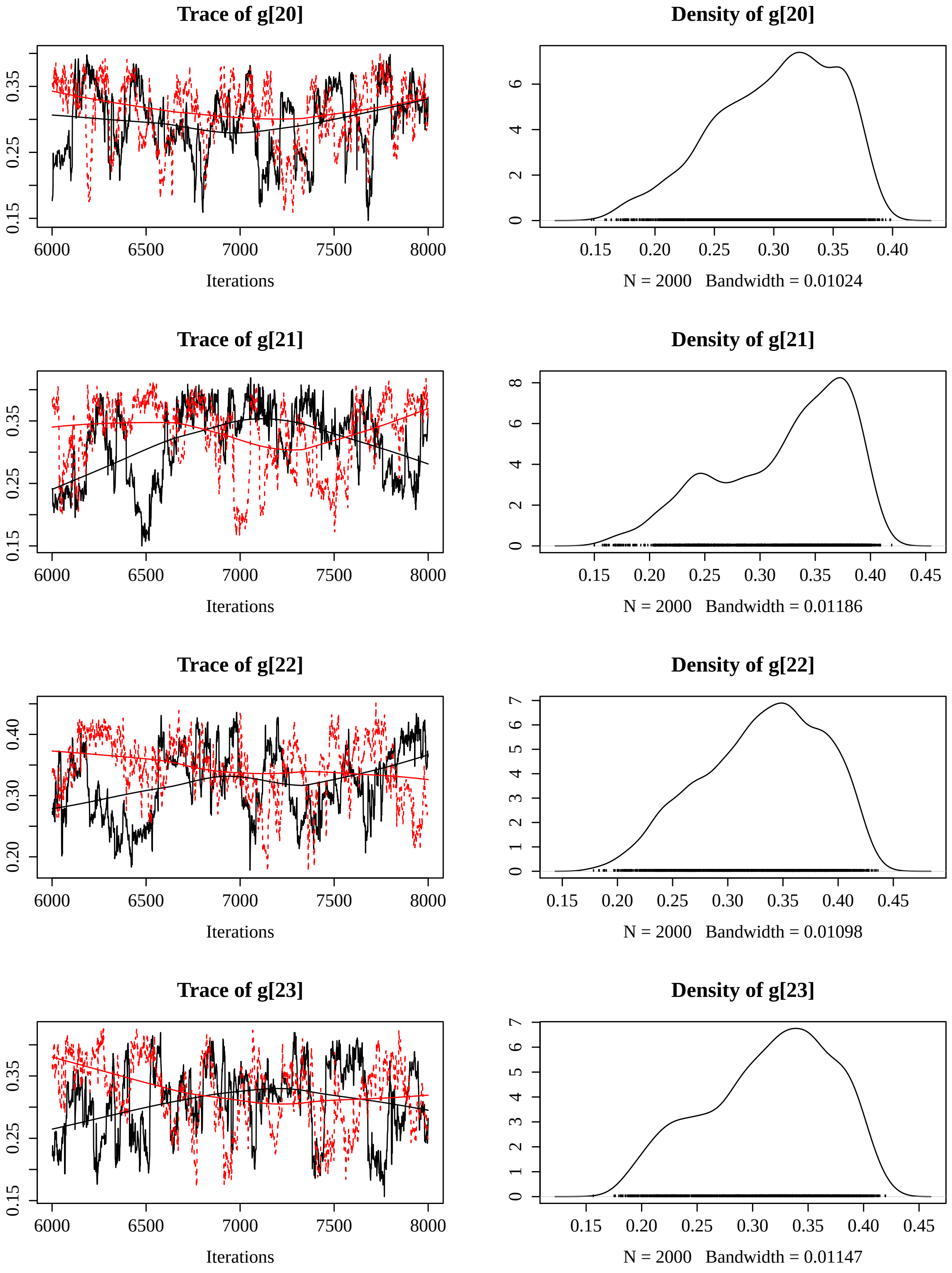


Figure B.3: Example Trace and Kernel Density Plots for c-parameter

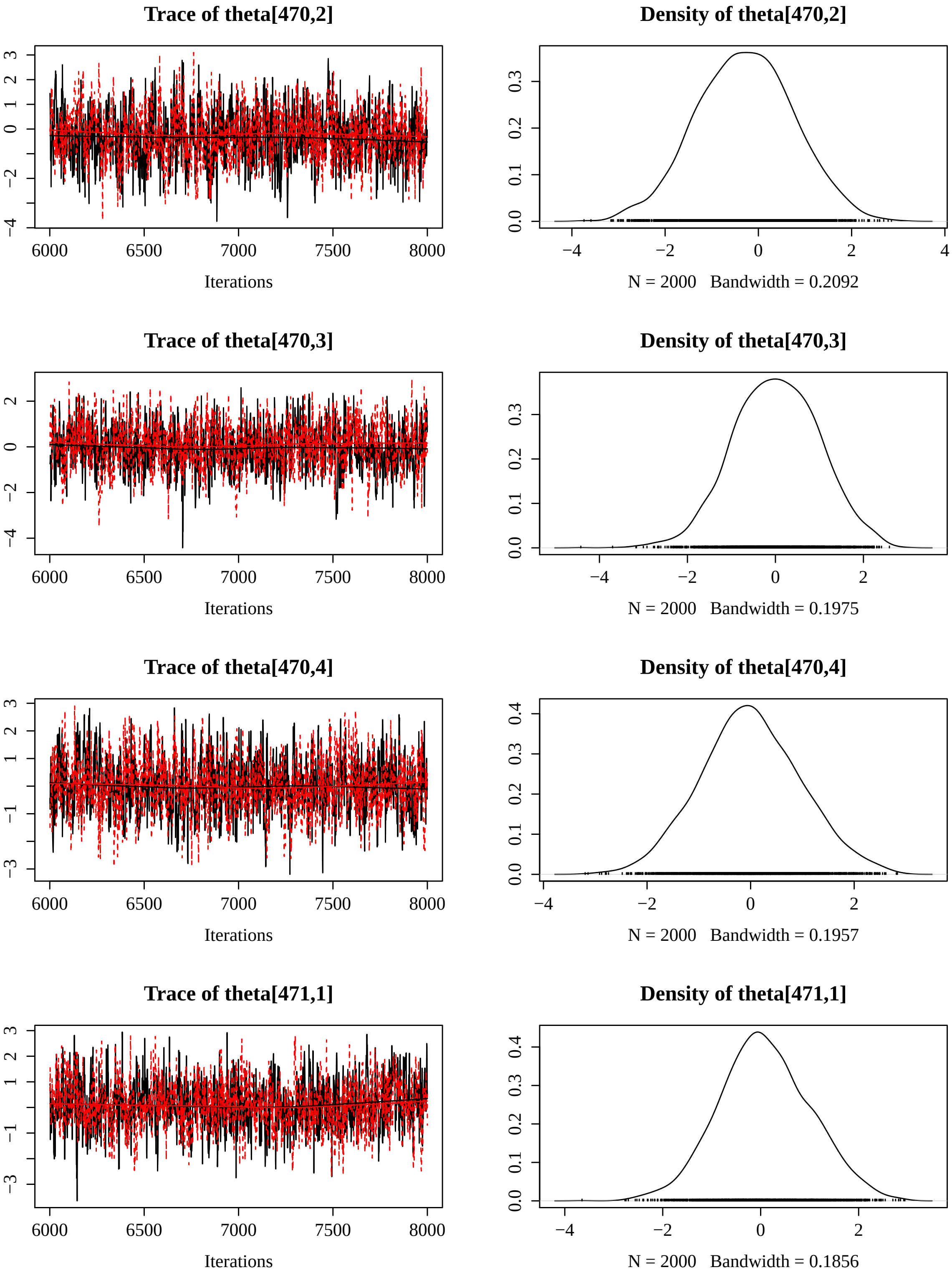


Figure B.4: Example Trace and Kernel Density Plots for Thetas

Table B.1: *Means and Standard Errors for Bias and RMSE of a-parameter*

Test Length	Discrimination Level	d, dimension	Bias		RMSE	
			M	*s.e.*	*M*	*s.e.*
30	0.50	1	0.593	0.040	0.638	0.042
		2	0.368	0.030	0.650	0.055
		3	0.364	0.030	0.643	0.044
		4	0.366	0.030	0.646	0.045
	0.75	1	0.604	0.040	0.649	0.041
		2	0.283	0.020	0.506	0.040
		3	0.281	0.030	0.503	0.046
		4	0.288	0.020	0.513	0.036
	1.00	1	0.579	0.040	0.624	0.034
		2	0.189	0.020	0.347	0.039
		3	0.197	0.020	0.364	0.040
		4	0.197	0.020	0.364	0.039
	1.25	1	0.600	0.040	0.644	0.040
		2	0.107	0.020	0.224	0.036
		3	0.110	0.020	0.228	0.030
		4	0.112	0.020	0.237	0.040
	1.50	1	0.609	0.040	0.653	0.042
		2	0.032	0.030	0.143	0.031
		3	0.030	0.020	0.134	0.024
		4	0.033	0.020	0.146	0.032
60	0.50	1	0.578	0.030	0.624	0.028
		2	0.358	0.020	0.632	0.027
		3	0.367	0.020	0.649	0.035
		4	0.358	0.020	0.633	0.028
	0.75	1	0.582	0.030	0.628	0.033
		2	0.277	0.010	0.497	0.023
		3	0.282	0.020	0.508	0.030
		4	0.272	0.020	0.488	0.027
	1.00	1	0.579	0.030	0.628	0.027
		2	0.191	0.010	0.355	0.026
		3	0.195	0.020	0.361	0.028
		4	0.194	0.020	0.363	0.028
	1.25	1	0.586	0.030	0.633	0.031
		2	0.111	0.020	0.235	0.029
		3	0.113	0.020	0.238	0.033
		4	0.108	0.020	0.227	0.025
	1.50	1	0.584	0.030	0.634	0.026
		2	0.029	0.020	0.146	0.024
		3	0.029	0.020	0.149	0.019
		4	0.028	0.020	0.134	0.023

Table B.2: *Means and Standard Errors for Bias and RMSE of b-parameter*

Test Length	Discrimination Level	d, dimension	Bias		RMSE	
			M	*s.e.*	*M*	*s.e.*
30	0.50	1	1.633	0.17	1.234	0.122
		2	0.545	0.16	1.095	0.135
		3	0.445	0.15	1.095	0.096
		4	0.569	0.16	1.131	0.150
	0.75	1	1.644	0.17	1.198	0.124
		2	0.544	0.15	1.106	0.135
		3	0.444	0.16	1.077	0.119
		4	0.574	0.16	1.129	0.144
	1.00	1	1.619	0.17	1.220	0.109
		2	0.533	0.15	1.097	0.131
		3	0.444	0.15	1.092	0.101
		4	0.567	0.15	1.130	0.131
	1.25	1	1.639	0.18	1.247	0.110
		2	0.535	0.16	1.102	0.139
		3	0.441	0.15	1.076	0.104
		4	0.565	0.16	1.129	0.135
	1.50	1	1.649	0.19	1.188	0.119
		2	0.543	0.15	1.088	0.128
		3	0.444	0.15	1.070	0.096
		4	0.57	0.16	1.127	0.143
60	0.50	1	1.593	0.13	1.215	0.090
		2	0.528	0.11	1.113	0.081
		3	0.531	0.11	1.096	0.091
		4	0.54	0.09	1.077	0.087
	0.75	1	1.595	0.13	1.220	0.079
		2	0.525	0.11	1.124	0.086
		3	0.53	0.11	1.098	0.090
		4	0.537	0.09	1.092	0.079
	1.00	1	1.592	0.12	1.231	0.083
		2	0.523	0.11	1.121	0.085
		3	0.526	0.11	1.088	0.086
		4	0.543	0.09	1.071	0.084
	1.25	1	1.601	0.13	1.217	0.095
		2	0.53	0.11	1.116	0.084
		3	0.528	0.1	1.069	0.076
		4	0.54	0.1	1.067	0.083
	1.50	1	1.599	0.13	1.222	0.080
		2	0.532	0.11	1.116	0.087
		3	0.527	0.11	1.072	0.078
		4	0.544	0.09	1.075	0.088

Table B.3: *Means and Standard Errors for Bias and RMSE of c-parameter*

Test Length	Discrimination Level	Bias		RMSE	
		M	*s.e.*	*M*	*s.e.*
30	0.50	0.120	0.005	0.015	0.001
	0.75	0.120	0.005	0.015	0.001
	1.00	0.120	0.005	0.015	0.001
	1.25	0.120	0.005	0.015	0.001
	1.50	0.120	0.005	0.015	0.001
60	0.50	0.121	0.003	0.015	0.001
	0.75	0.120	0.003	0.015	0.001
	1.00	0.121	0.003	0.015	0.001
	1.25	0.120	0.003	0.015	0.001
	1.50	0.120	0.003	0.015	0.001

Table B.4: *Means and Standard Errors for Bias, RMSE, and Pearson's Correlation for Thetas*

TL	DL	*d*	Bias		RMSE		Correlation	
			M	*s.e.*	*M*	*s.e.*	*M*	*s.e.*
30	0.50	1	-0.01	0.001	0.976	0.004	0.98	0.006
		2	0.018	0.001	1.012	0.002	0.996	0.003
		3	-0.036	0.001	1.01	0.003	0.997	0.002
		4	-0.033	0.001	0.995	0.003	0.996	0.003
	0.75	1	-0.01	0.001	0.977	0.004	0.981	0.007
		2	0.018	0.001	1.01	0.003	0.994	0.004
		3	-0.036	0.001	1.008	0.002	0.995	0.003
		4	-0.033	0.001	0.993	0.003	0.994	0.004
	1.00	1	-0.01	0.001	0.977	0.004	0.981	0.007
		2	0.018	0.001	1.008	0.003	0.991	0.005
		3	-0.036	0.001	1.006	0.003	0.992	0.004
		4	-0.033	0.001	0.991	0.003	0.99	0.005
	1.25	1	-0.01	0.001	0.977	0.004	0.981	0.006
		2	0.018	0.001	1.008	0.003	0.99	0.005
		3	-0.036	0.001	1.005	0.003	0.99	0.004
		4	-0.033	0.001	0.991	0.003	0.989	0.005
	1.50	1	-0.01	0.001	0.977	0.004	0.981	0.006
		2	0.018	0.001	1.006	0.003	0.987	0.007
		3	-0.036	0.001	1.004	0.003	0.988	0.005
		4	-0.033	0.001	0.989	0.003	0.987	0.006
60	0.50	1	-0.014	0.001	0.963	0.005	0.954	0.008
		2	0.017	0.001	1.012	0.004	0.993	0.004
		3	-0.037	0.001	1.012	0.004	0.996	0.003
		4	-0.035	0.001	0.995	0.003	0.993	0.004
	0.75	1	-0.015	0.001	0.964	0.005	0.956	0.008
		2	0.017	0.001	1.008	0.004	0.988	0.005
		3	-0.037	0.001	1.007	0.003	0.992	0.004
		4	-0.036	0.001	0.992	0.003	0.989	0.005
	1.00	1	-0.014	0.001	0.965	0.005	0.957	0.008
		2	0.017	0.001	1.004	0.005	0.983	0.008
		3	-0.037	0.001	1.004	0.004	0.988	0.006
		4	-0.035	0.001	0.989	0.003	0.985	0.006
	1.25	1	-0.015	0.001	0.965	0.005	0.957	0.008
		2	0.016	0.001	1.002	0.004	0.979	0.008
		3	-0.037	0.001	1.002	0.004	0.984	0.006
		4	-0.035	0.001	0.986	0.003	0.98	0.006
	1.50	1	-0.014	0.001	0.966	0.005	0.958	0.007
		2	0.016	0.001	1	0.005	0.975	0.01
		3	-0.037	0.001	1.001	0.004	0.982	0.007
		4	-0.035	0.001	0.984	0.004	0.977	0.008

Note: TL: Test Length, DL: Discrimination Level, *d*: dimension