

## ABSTRACT

Title of Dissertation: IRT VS. FACTOR ANALYSIS APPROACHES  
IN ANALYZING MULTIGROUP  
MULTIDIMENSIONAL BINARY DATA: THE  
EFFECT OF STRUCTURAL  
ORTHOGONALITY, AND THE  
EQUIVALENCE IN TEST STRUCTURE, ITEM  
DIFFICULTY, & EXAMINEE GROUPS

Peng Lin, Doctor of Philosophy, 2008

Directed By: Professor Robert W. Lissitz  
Department of Measurement, Statistics and  
Evaluation

The purpose of this study was to investigate the performance of different approaches in analyzing multigroup multidimensional binary data under different conditions. Two multidimensional Item Response Theory (MIRT) methods (concurrent MIRT calibration and separate MIRT calibration with linking) and one factor analysis method (concurrent factor analysis calibration) were examined. The performance of the unidimensional IRT method compared to its multidimensional counterparts was also investigated.

The study was based on simulated data. Common-item nonequivalent groups design was employed with the manipulation of four factors: the structural orthogonality, the equivalence of test structure, the equivalence of item difficulty, and the equivalence of examinee groups. The performance of the methods was evaluated based on the recovery of the item parameters and the estimation of the true score of the examinees.

The results indicated that, in general, the concurrent factor analysis method performed as well as, sometimes even better than, the two MIRT methods in recovering the item parameters. However, in estimating the true score of examinees,

the concurrent MIRT method usually performed better than the concurrent factor analysis method. The results also indicated that the unidimensional IRT method was quite robust to the violation of unidimensionality assumption.

IRT vs. Factor Analysis Approaches in Analyzing  
Multigroup Multidimensional Binary Data:  
The Effect of Structural Orthogonality, and the Equivalence in Test Structure, Item  
Difficulty, & Examinee Groups

By

Peng Lin

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College park, in partial fulfillment  
Of the requirements for the degree of  
Doctor of Philosophy  
2008

Advisor Committee:  
Professor Robert W. Lissitz, Chair  
Professor Gregory R. Hancock  
Professor Robert J. Mislevy  
Assistant Professor Jeffrey R. Harring  
Professor Edward L. Fink

## DEDICATION

To my family, my beloved ones.

## ACKNOWLEDGEMENTS

I would like to give my deepest gratitude to Dr. Lissitz, my dissertation advisor, for his guidance, patience, and encouragement on my dissertation and many other works. He has been a great mentor for me not only in graduate school but also in life.

My true thanks go to Dr. Hancock, my academic advisor, for his directions on my academic works throughout my doctoral study, and for his suggestions and all kinds of supports on my dissertation work. I sincerely appreciate Dr. Mislevy, Dr. Fink, and Dr. Haring for their time and comments on my dissertation.

Special thanks go to my husband, Guojing, for his support, encouragement, and love; to my sweet daughter, Claire, for being a healthy and happy baby and having brought me the most amazing experience in my life; to my parents and sister, for always being so supportive to me in countless ways.

## TABLE OF CONTENTS

LIST OF TABLES .....	vi
LIST OF FIGURES .....	vii
CHAPTER 1 INTRODUCTION .....	1
1.1 Background and Research Questions.....	1
1.2 Group Invariance and Scale Indeterminacy in Multigroup Analysis.....	5
1.3 Multigroup Unidimensional Analysis .....	8
1.3.1 Unidimensional IRT (UIRT) methods .....	8
1.3.2 Uni-factor Analysis Models .....	9
1.4 Multigroup Multidimensional Analysis .....	11
1.4.1 Multidimensional IRT (MIRT) models .....	11
1.4.2 Multi-factor analysis models.....	13
1.5 Purpose of the Study .....	14
CHAPTER 2 LITERATURE REVIEW.....	15
2.1. IRT Models .....	15
2.1.1 Unidimensional IRT (UIRT) Model and Linking .....	15
2.1.2 The Multidimensional IRT Model .....	16
2.1.3 Concurrent and separate MIRT calibration methods .....	19
2.1.4 Linking in MIRT .....	22
2.2 Factor analysis models.....	26
2.2.1 Factor analysis models for continuous data.....	26
2.2.2 Factor analysis models for categorical data.....	27
2.3. The equivalence of IRT and factor analysis method.....	30
2.3.1 Normal-ogive IRT model vs. logistic IRT model.....	30
2.3.2 The relationship between normal-ogive IRT model and factor analysis model.....	31
CHAPTER 3 METHODOLOGY .....	33
3.1 Simulation Design.....	33
3.2 The Multigroup Analysis Methods Investigated.....	34
3.3 Key factors .....	36
3.4 Data generation .....	41
3.4.1 Item parameters generation.....	41
3.4.2. Generation of Correlated $\theta_1$ and $\theta_2$ for Group 1 and Group 2.....	43
3.4.3 Generate Response Data .....	44
3.5 Linking for Separate MIRT Calibration.....	45
3.6 Evaluation Criteria.....	46
CHAPTER 4 RESULTS .....	49
4.1 Recovery of the item parameters .....	51
4.1.1 The recovery of $a_1$ .....	52
4.1.2 The recovery of $a_2$ .....	61
4.1.3 The recovery of $d$ .....	70
4.2 Estimate of true score of the examinees .....	77
4.2.1 The estimate of true scores in Group 1 .....	77
4.2.2 The true score estimation in Group 2.....	85

CHAPTER 5 CONCLUSION AND DISCUSSION.....	93
5.1 Summary of the Study .....	93
5.2 Summary of the Results .....	95
5.2.1 The Recovery of Item Parameters.....	95
5.2.2 The Estimation of True Score of Examinees .....	98
5.3 Discussion and Future Study .....	100
APPENDIX A .....	104
APPENDIX B .....	106
APPENDIX C .....	112
REFERENCES .....	116

## LIST OF TABLES

Table 3-1 The number of items measuring or or both.....	38
Table 3-2 The mean proficiency on and in the two groups.....	39
Table 4-1 The 54 combinations of conditions.....	50
Table B-1 <i>BIAS</i> of $a_1$ estimated from the three methods under all 54 conditions .....	106
Table B-2 <i>SD</i> of $a_1$ estimated from the three methods under all 54 conditions .	107
Table B-3 <i>BIAS</i> for $a_2$ estimated from the three methods under all 54 conditions .....	108
Table B-4 <i>SD</i> for $a_2$ estimated from the three methods under all 54 conditions .....	109
Table B-5 <i>BIAS</i> for $d$ estimated from the three methods under all 54 conditions .....	110
Table B-6 <i>SD</i> for $d$ estimated from the three methods under all 54 conditions	111
Table C-1 <i>BIAS</i> of true score estimated from the three methods .....	112
Table C-2 <i>SD</i> of true score estimated from the three methods.....	113
Table C-3 <i>BIAS</i> of true score estimated from the three methods .....	114
Table C-4 <i>SD</i> of true score estimated from the three methods.....	115



## LIST OF FIGURES

Figure 1-1 Content Specification in a Grade 3-8 Mathematics Assessment Blueprint .....	3
Figure 1-2 Linking in (a)unidimensional and (b)multidimensional models.....	7
Figure 1-3 Multigroup unidimensional factor analysis.....	10
Figure 1-4 Concurrent estimation in multigroup multidimensional SEM analysis .....	13
Figure 2-2 The item vectors in the ability space.....	19
Figure 2-3 The categorization of the continuous into dichotomous .....	28
Figure 3-1 Items in Form 1 and Form 2.....	34
Figure 3-2 The distribution of the items of Form 1 in the ability space .....	37
Figure 3-3 Generating correlated $\theta_1$ and $\theta_2$ from higher order variable $z$ .....	44
Figure 4-1 <i>BIAS</i> of $a_1$ from the three calibration methods under all 54 conditions .....	55
Figure 4-2 <i>SD</i> of $a_1$ from the three calibration methods under all 54 conditions	56
Figure 4-3 <i>BIAS</i> and <i>SD</i> of $a_1$ under three structural orthogonality levels .....	57
Figure 4-4 <i>BIAS</i> and <i>SD</i> of $a_1$ under three structural equivalence levels.....	58
Figure 4-5 <i>BIAS</i> and <i>SD</i> of $a_1$ under two item difficulty equivalence levels.....	59
Figure 4-6 <i>BIAS</i> and <i>SD</i> of $a_1$ under three examinee group equivalence levels .	60
Figure 4-7 <i>BIAS</i> of $a_2$ from the three calibration methods under all 54 conditions.....	64
Figure 4-8 <i>SD</i> of $a_2$ from the three calibration methods under all 54 conditions .....	65
Figure 4-9 <i>BIAS</i> and <i>SD</i> of $a_2$ under three structural orthogonality levels.....	66
Figure 4-10 <i>BIAS</i> and <i>SD</i> of $a_2$ under three structural equivalence levels .....	67
Figure 4-11 <i>BIAS</i> and <i>SD</i> of $a_2$ under two item difficulty equivalence levels ...	68
Figure 4-12 <i>BIAS</i> and <i>SD</i> of $a_2$ under three examinee group equivalence levels .....	69

Figure 4-13 <i>BIAS</i> of $d$ from the three calibration methods under all 54 conditions.....	71
Figure 4-14 <i>SD</i> of $d$ from the three calibration methods under all 54 conditions .....	72
Figure 4-15 <i>BIAS</i> and <i>SD</i> of $d$ under three structural orthogonality levels.....	73
Figure 4-16 <i>BIAS</i> and <i>SD</i> of $d$ under three structural equivalence levels.....	74
Figure 4-17 <i>BIAS</i> and <i>SD</i> of $d$ under two item difficulty equivalence levels ....	75
Figure 4-18 <i>BIAS</i> and <i>SD</i> of $d$ under three examinee group equivalence levels	76
Figure 4-19 <i>BIAS</i> of the true score estimation in Group 1.....	79
Figure 4-20 <i>SD</i> of the true score estimation in Group 1.....	80
Figure 4-21 <i>BIAS</i> and <i>SD</i> of estimation of true score in Group 1 under three structural orthogonality levels.....	81
Figure 4-22 <i>BIAS</i> and <i>SD</i> of estimation of true score in Group 1 under three structural equivalence levels .....	82
Figure 4-23 <i>BIAS</i> and <i>SD</i> of estimation of true score in Group 1 under two item difficulty equivalence levels .....	83
Figure 4-24 <i>BIAS</i> and <i>SD</i> of estimation of true score in Group 1 under three examinee group equivalence levels .....	84
Figure 4-25 <i>BIAS</i> of the true score estimation in Group 2.....	87
Figure 4-26 <i>SD</i> of the true score estimation in Group 2.....	88
Figure 4-27 <i>BIAS</i> and <i>SD</i> of estimation of true score in Group 2 under three structural orthogonality levels.....	89
Figure 4-28 <i>BIAS</i> and <i>SD</i> of estimation of true score in Group 2 under three structural equivalence levels .....	90
Figure 4-29 <i>BIAS</i> and <i>SD</i> of estimation of true score in Group 2 under two item difficulty equivalence levels .....	91
Figure 4-30 <i>BIAS</i> and <i>SD</i> of estimation of true score in Group 2 under three examinee group equivalence levels .....	92

# CHAPTER 1

## INTRODUCTION

### 1.1 Background and Research Questions

In educational assessments, multigroup analysis has been widely applied in equating, vertical scaling, differential item functioning (DIF) analysis, and two-stage testing. According to Bock and Zimowski (1996):

[Multigroup analysis provides] . . . a unified approach to such problems as *differential item functioning, item parameter drift, nonequivalent groups equating, vertical equating, two-stage testing, and matrix-sampled educational assessment*. The common element in these problems is the existence of persons from different populations responding to the same test or to tests containing common items . . . , the objective of the multiple-group analysis is to estimate jointly the item parameters and the latent distribution of a common attribute or ability of the persons in each of the populations (Bock & Zimowski, 1996, p. 433).

In practice, most of the approaches applied in the multigroup analysis are unidimensional. However, a limitation of the unidimensional approaches is that the assumption of unidimensionality sometimes does not hold, even though the statistical analysis proves it acceptable. Most, if not all, tests measure a complex of abilities rather than a single one (Reckase, Ackerman, & Carlson, 1988). For example, a state mathematics accountability test might be developed to measure abilities on algebra, geometry, data and probability, measurement, and number and operations. Although these abilities might be related, the relationship would hardly be perfect. For example, two examinees with equal ability on geometry might have different abilities on algebra. When only one item is of interest, a unidimensional model can always work well because the resulting single dimension might represent a single ability or a composite of abilities. However, when a set of items are considered, the use of unidimensional models must be considered carefully (Ackerman, 1994). A variety of

research has been conducted to investigate the robustness of the unidimensional models to multidimensional data when only one group or test is considered. Reckase and Ackerman (1988) stated that when the same weighted composite of multiple abilities is measured by all items of a test, the test can be treated as unidimensional. Wang (1986) and Dickenson (2005) showed that when using a unidimensional IRT model analyzing multidimensional data, the resulting single dimension is actually a linear composite of the multiple dimensions. Min (2003) summarized three different conditions under which applying unidimensional models is appropriate: 1) both examinee's ability and test item characteristics are varying on one dimension as assumed in the model; 2) examinee ability varies only on one ability dimension even though test items are measuring more than one ability; 3) examinee abilities are different on multiple ability dimensions but all items are measuring the same composite of abilities. In other conditions that cannot be categorized as one of the three above, applying unidimensional models might be problematic. Studies have shown that when the multidimensional data are modeled under the unidimensional assumption, measurement error will increase and the inferences from the results would be problematic (Ackerman, 1994; Baker, 1992; Reckase, 1985, 1995).

In multigroup analysis, the application of unidimensional models should be considered even more carefully than in the single group analysis because not only the structure of the test in each group could be multidimensional, but the dimensions in the test structure could change across groups. Again, use the state mathematics accountability test as an example to illustrate this situation, as is shown in Figure 1-1. The figure describes the content specification of the test in Grades 3 through 8. In Grade 3, the test is developed to measure geometry, algebra, number and operation. From Grade 4, one additional ability, data and probability, is added in the test. In

Grade 7 and Grade 8, another additional ability, measurement, is tested. However, in Grade 8, algebra is no longer tested. From this example, we can see that the test at each grade measures multiple abilities and the abilities are not consistent across grades. Even when the test at Grades 4 to 6 measure the same four abilities, the measurement emphases on these four abilities are not same for different grades. Under this condition, using a unidimensional model might not be able to capture the changes in the test structure, and therefore, the illustration of the test results based on the unidimensional model would be problematic.

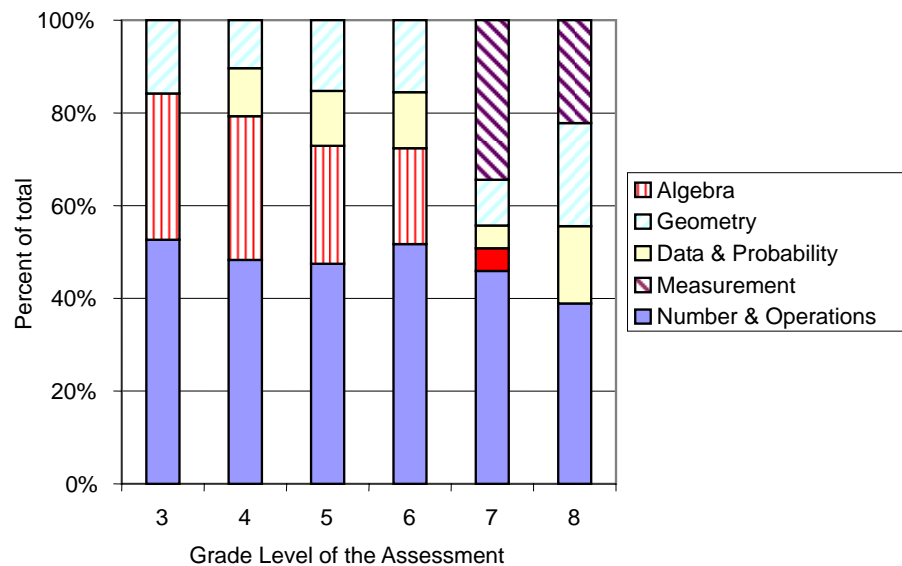


Figure 1-1 Content Specification in a Grade 3-8 Mathematics Assessment Blueprint  
(Martineau, 2006)

To solve this problem, multidimensional approaches have been proposed for multigroup analysis. Multidimensional Item Response Theory (IRT) methods and factor analysis methods are two important ones. IRT methods have been widely applied in analyzing the tests where the items are scored dichotomously (0 vs. 1) or polytomously (e.g., 0, 1, 2) (Kolen & Brennan, 2005). Although factor analysis

models are often applied in the situations in which the indicators are treated as continuous (e.g., the total score of a test or the score of a testlet), they can also be applied to analyze categorical, dichotomous or polytomous, item response data (Bock & Aitkin, 1981; Chirstoffersson, 1975; Horst, 1965; McDonald, 1967; Muthén, 1978). Both IRT methods and factor analysis methods provide powerful tools to describe the relationship between item responses and the latent traits, as well as estimate the relative amount of the latent traits of the examinees. If multiple traits are measured by the test, some factor analysis models (Structural Equation Modeling models) can also describe the causal relationship between the latent traits. IRT methods, however, do not provide this kind of information. The discussion of causal relationship between latent traits is beyond the scope of this study. Although IRT and factor analysis methods belong to different traditions, they are highly related (Glockner-Rist & Hoijtink, 2003; Knol & Berger, 1991; Reckase, 1997; Takane & de Leeuw, 1987). According to Takane and de Leeuw (1987), when the latent traits are normally distributed, IRT and factor analysis models are equivalent. The performance of IRT and factor analysis methods in single group analysis has been investigated in previous literature (Glockner-Rist & Hoijtink, 2003; Knol & Berger, 1991). But there have been few studies that investigate the performance of the two methods in multigroup analysis.

How do IRT methods and factor analysis methods perform in multigroup analysis? How is the performance of these methods affected by the characteristics of the tests? Do multidimensional methods have evident advantages over the unidimensional counterparts in analyzing multidimensional data? These are the questions this study explores.

## 1.2 Group Invariance and Scale Indeterminacy in Multigroup Analysis

One important assumption of multigroup analysis is measurement invariance.

That is, the parameters of any given item are the same for all groups. Rupp and

Zumbo (2006) stated that:

. . . for inferences to be equally valid for different populations of examinees or different measurement conditions, parameters in the psychometric models used for data analysis need to be invariant; if parameters are not invariant, the statistical foundation for inferences is not identical across the populations or measurement conditions, and hence the inferences are not generalizable across those to the same degree (Rupp & Zumbo, 1996, p. 64).

If the assumption of measurement invariance is violated, it might indicate the presence of differential item functioning (DIF) (the parameters of a given item are different across the groups formed by gender or other demographic features), or item parameter drift (IPD) (the parameters of a given item change over subsequent occasions) (Goldstein, 1983).

When the parameters are estimated separately for each group, the estimate of the parameters of the same item might be different across groups. However, one cannot simply conclude that measurement invariance does not hold because scale difference (using a different scale measuring the parameters in different groups) can also lead to such discrepancy. A frequently cited example of this situation is measuring temperature using different scales. Assume one person uses the Fahrenheit scale and reads the temperature as 32°, whereas another one uses the Celsius scale and reads the temperature as 0°. The difference between the two reads does not indicate that the temperature is different. It is just a result of scale difference.

In IRT and factor analysis models, the scale of item or person parameters is quite arbitrary. In the unidimensional models, the origin and unit can be set at any value without changing the fit of the model. This is often referred to as *scale indeterminacy*: the scale of parameters is determined only up to a linear

transformation (Oshima et al., 2000). In most cases, a scale is selected so that the mean and standard deviation of the latent traits are 0 and 1 (Zimowski, 2003), which is called *standardization*. In multigroup analysis, when the distribution of the latent traits is not equivalent across groups, standardization within each group might result in different scales for different groups. Under this condition, the parameters estimated from different groups can not be compared directly. What's more important, the inference made based on the parameter estimates in one group may not be generalized to other groups. The scale indeterminacy problem in multidimensional models is more complicated than that in unidimensional models. In addition to the indeterminacy of origin and scale, multidimensional models have an additional indeterminacy, rotation indeterminacy, the direction the dimensions can be rotated in the ability space without changing the model fit (Li & Lissitz, 2000; Min, 2003; Reckase & Martineau, 2004).

To solve the problems caused by scale difference, a common scale for all groups is needed. One approach to achieving a common scale is to estimate the parameters simultaneously for all groups and constrain the parameters of the same item to be equal across groups. This method is often referred to as the *concurrent calibration*. Another approach is to estimate the parameters separately for each group and then rescale the parameters onto the common scale. This method is often referred to as the *separate calibration* and the process of rescaling is often referred to as *linking*. Figure 1-2 (from Min, 2003, with some changes) illustrates the scale transformation in (a) unidimensional and (b) multidimensional models. In the figure, Scale  $B$  is used as the common scale or base scale to which Scale  $E$  is transformed.  $O_B$  is the origin for Scale  $B$  and  $O_E$  is the origin for Scale  $E$ . The unit for Scale  $B$  is the segment between point  $O_B$  and  $U_B$ , and that for Scale  $E$  is the segment between point  $O_E$  and  $U_E$ . During scale transformation, the origin of Scale  $E$  is



shifted to Scale  $B$  by translation, the unit of Scale  $E$  is adjusted to Scale  $B$  by dilation, and for multidimensional linking, the coordinate system of Scale  $E$  is aligned with that of Scale  $B$  through rotation.

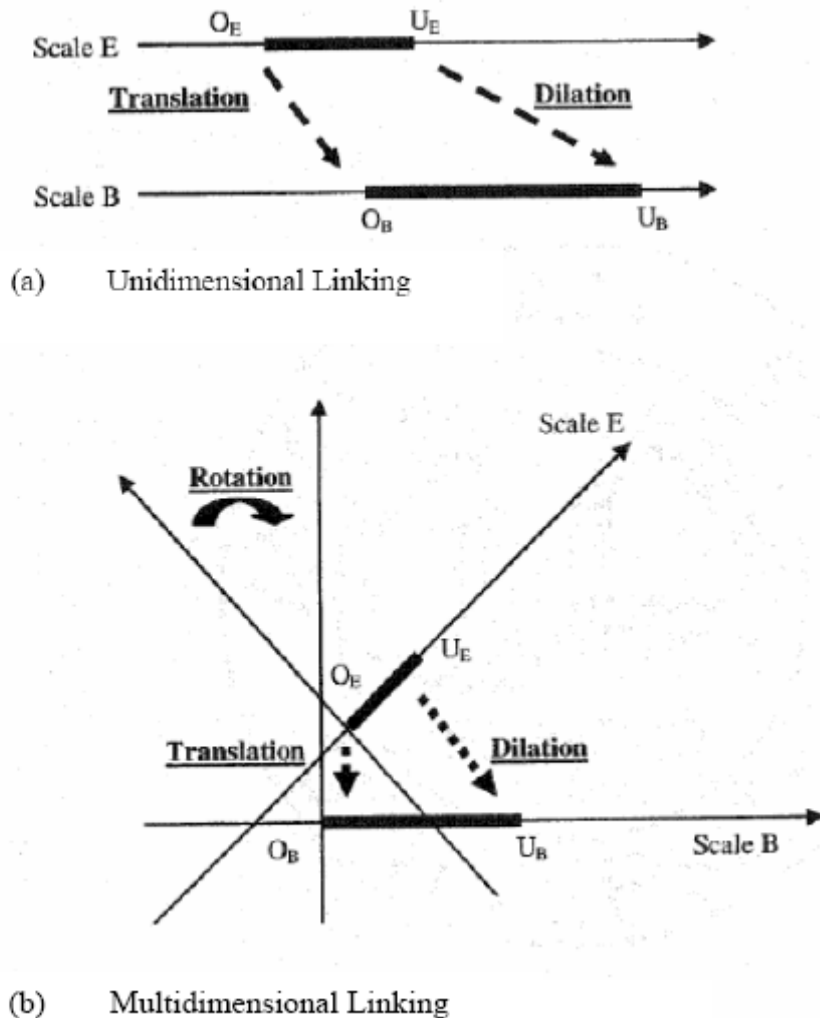


Figure 1-2 Linking in (a) unidimensional and (b) multidimensional models (resource: Min, 2003)

Both concurrent calibration and separate calibration have their merits and limits. One prominent advantage of concurrent calibration is that it estimates the parameters for all groups at one time, and there is no need for linking (Kolen & Brennan, 2004). Studies also indicated that for unidimensional models, concurrent

calibration produces less biased and more stable estimate than separate calibration when the data fit the model (Hanson & Béguin, 2002; Kim, 2004; Kim & Cohen, 1998; Spence, 1996; Yao & Mao, 2004). However, one limit of concurrent calibration is that it has a higher requirement on both the computer program and computer capacity than separate calibration.

When the parameter estimates are put on a common scale, further analysis can be conducted. For example, scores from parallel forms of a test can be equated (equating), or the growth of the examinees can be evaluated through a battery of tests scanning several years.

### 1.3 Multigroup Unidimensional Analysis

#### 1.3.1 Unidimensional IRT (UIRT) methods

There are a variety of unidimensional IRT models. For example, the models for binary (dichotomous) data include Rasch model, two-parameter logistic model (2-PLM), three-parameter logistic model (3-PLM), and Normal-ogive model. The models for polytomous data include partial credit model and graded response model.

Mislevy (1987) and Bock and Zimowski (1996) described the multigroup IRT procedures for concurrently estimating item and ability parameters for all groups using the maximum marginal likelihood (MML) method (Bock & Aitkin, 1981; Bock & Lieberman, 1970). During the process of estimation, the item parameters are estimated over all groups whereas the ability distribution is estimated separately for each group so that they can be different when the groups are nonequivalent. The procedures have been incorporated in the computer program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) for dichotomous data and in PARSCALE (Muraki & Bock, 1991) and PARDUX (Burket, 2002) for polytomous data.

Several programs have been developed for separate calibration, such as LOGIST (Wingersky, Barton, Lord, 1982) and BILOG 3 (Mislevy & Bock, 1990). When the parameters are estimated separately for each group, the estimates from different groups need to be linked. Usually, one group is selected as the reference group and the scale of parameters in the reference group is treated as the base scale, onto which the parameter estimated from other groups are transformed through some transformation equations.

A variety of studies have been conducted to compare unidimensional concurrent and separate calibration (Béguin & Hanson, 2001; Béguin, Hanson, & Glas, 2000; Hanson & Béguin, 2002; Kim, 2004; Kim & Cohen, 1998; Spence, 1996). These studies suggested that when the data fit the IRT model, concurrent calibration produced less biased and more stable estimate than separate calibration. However, when the data violate the assumption of unidimensionality, the advantage of concurrent estimation is questionable. Some studies (Béguin & Hanson, 2001; Béguin, Hanson, & Glas, 2001; Yao & Mao, 2004) indicated that, while doing equating, the separate calibration might be more robust to multidimensionality than the concurrent calibration. However, some other studies (Kim, 2004; Spence, 1996) came to the opposite conclusion.

### 1.3.2 Uni-factor Analysis Models

The unidimensional normal-ogive IRT model (Bock & Lieberman, 1970) and the general multigroup factor analysis methods for continuous variables (Jöreskog, 1971; Sörbom, 1974) are two origins of multigroup uni-factor analysis models. Multigroup factor analysis methods for continuous data concurrently estimate the parameters by constraining the parameters (factor loadings and thresholds) of the

same item to be equal across groups and allow the ability distribution to be different for nonequivalent groups, which is illustrated in Figure 1-3, where  $\theta$  is the ability measured by the items; the double-headed arched arrow represents the variance of  $\theta$ ; the one-headed arrows from  $\theta$  to the items represent the factor loadings, which depict the relationship between the item responses and the latent trait; and the equal sign indicates that the parameters of given common item are constrained to be equal across groups during the process of estimation. For categorical data, factor analysis methods assume that there is a latent continuous variable underlying each categorical variable. The categorical data are formed by categorizing the latent continuous variable based on the threshold(s). In multigroup analysis, the threshold(s) for each common item should be equal across groups. The details of this method are discussed in the next chapter. The multigroup factor analysis can be carried out in the computer programs such as LISREL (Jöreskog & Sörbom, 2004), Mplus (Muthén & Muthén, 2006), and EQS (Bentler, 2004).

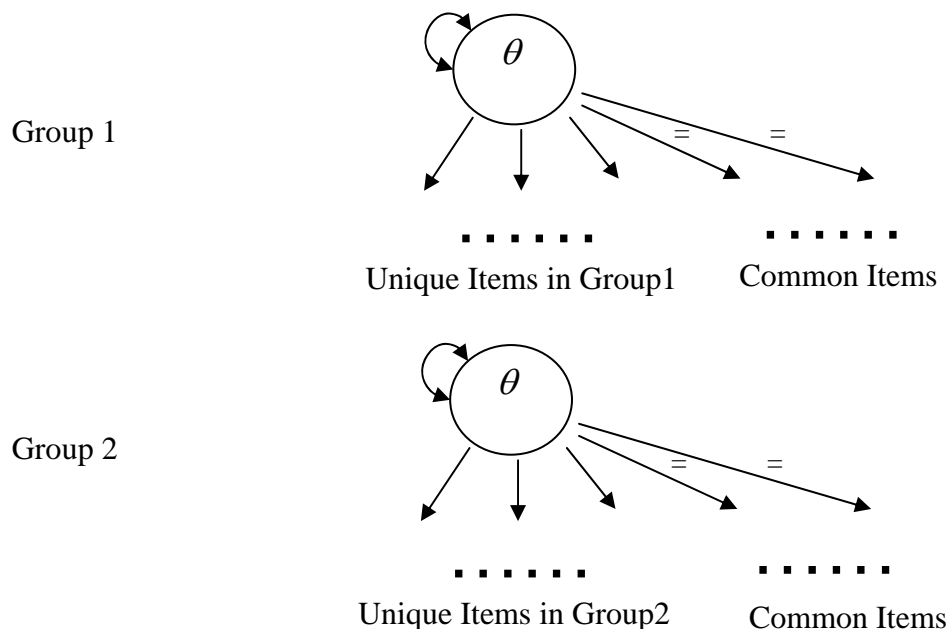


Figure 1-3 Multigroup unidimensional factor analysis

In educational assessment, factor analysis methods are less frequently applied than IRT methods in unidimensional analysis, perhaps because, in practice, one of the main purposes of using factor analysis methods is to explore the dimensionality structure of the items, whereas that of using IRT methods is to explore the interaction between item response and the latent trait.

#### 1.4 Multigroup Multidimensional Analysis

##### 1.4.1 Multidimensional IRT (MIRT) models

Most of the MIRT models are derived by generalizing their unidimensional counterparts to multidimensional models. The examples include the multidimensional 2-PLM, multidimensional 3-PLM, and multidimensional Normal-ogive model, for binary (dichotomous) data; multidimensional partial credit model, and multidimensional graded response model for polytomous data.

Concurrent calibration of the MIRT models in multigroup analysis can be carried out by a Bayesian based approach proposed by Yao (2003, 2004) which employs Markov Chain Monte Carlo (MCMC) methods to estimate the parameters. The procedure has been implemented in a computer program BMIRT (Yao, 2003).

In separate calibration, the parameters of MIRT models are first estimated for each group by computer programs, such as TESTFACT or NOHARM (note that BMIRT can also do separate calibration). Then the parameters estimated from different groups are linked to a common scale. Several approaches have been proposed for MIRT scale linking. Davey and his colleagues (Davey, Oshima, & Lee, 1996; Oshima, Davey, & Lee, 2000) proposed four procedures for multidimensional scale linking, three of which were implemented in computer program IPLINK (Lee & Oshima, 1996). Davey et al. 's (1996, 2000) methods allow oblique rotation of the

latent structure so that the rotation matrix is supposed to adjust both the unit and the orientation of the dimensions. Li and Lissitz (2000) proposed three transformation procedures from a different perspective than the Davey et al.'s (1996, 2000) methods. In their procedures, it is assumed that the dimensions are orthogonal (not assumed in Davey et al.'s procedures), and only orthogonal rotation is allowed. In Li and Lissitz's (2000) approach the work of rotation in Davey et al.'s (1996, 2000) procedures is split into two parts, where the Procrustes orthogonal rotation matrix adjusts the orientation of the dimensions and a dilation scalar adjust the unit. The three procedures were implemented in the program MDEQUATE (Li, 1996). Min (2003) extended Li and Lissitz's (2000) approach by allowing the unit dilation to be different for different dimensions. Min's (2003) approach works well when the number of dimension is low, but when the number of dimensions is high, the computational burden becomes unfeasible (Reckase & Martineau, 2004). To address this flaw, Reckase and Martineau (2004) employed a non-orthogonal Procrustes transformation approach (Mulaik, 1972), which automatically aligns each dimension of the original matrix to the target matrix (the base matrix) without assuming orthogonality. This approach eliminates the need for a dilation parameter without causing a scale indeterminacy problem. From this point, Reckase and Martineau's (2004) procedure is similar to Davey et al.'s (1996, 2000), although they use somewhat different methods to determine the transformation equation. Yon and Reckase (2005) compared Davey et al.'s (1996, 2000) procedure and Reckase and Martineau's (2004) procedure in the performance of MIRT parameter recovery in multigroup analysis. They found that for the mixed structure data (the item measures more than one dimensions) Reckase and Martineau's (2004) non-orthogonal Procrustes procedure works to some degree better than Davey et al.'s (1996, 2000).

### 1.4.2 Multi-factor analysis models

Present in literature, most of the factor analysis models are multi-factor ones. As has been discussed earlier, multi-factor analysis can not only explore the interaction between the observed data and the latent traits, but also provide more flexibilities than IRT models in analyzing the connection among the latent traits.

The procedure employed by multi-factor analysis models in multigroup analysis is similar to that used in uni-factor analysis. Figure 1-4 illustrates the concurrent estimation of parameters in multigroup multidimensional factor analysis. The equal signs indicate that the factor loading for the same item is constrained to be equal across groups. The programs such as LISREL (Jöreskog & Sörbom, 2004), Mplus (Muthén & Muthén, 2006), and EQS (Bentler, 2004) can also conduct multigroup multidimensional analysis.

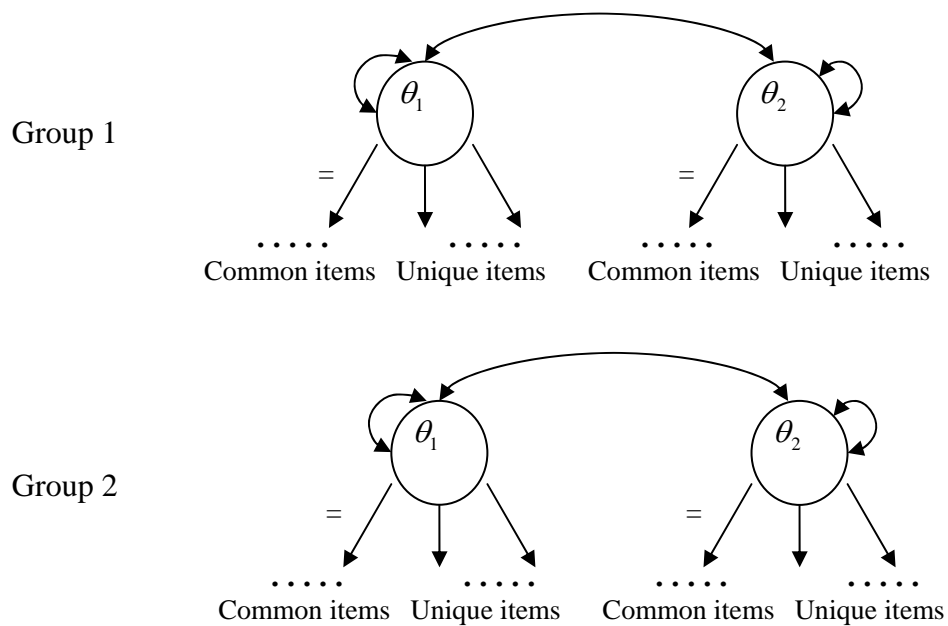


Figure 1-4 Concurrent estimation in multigroup multidimensional SEM analysis

Most of the multigroup factor analysis studies conducted previously assumed that same set of items are given to different groups (Jöreskog, 1971; Sörbom, 1974;

Muthén & Christoffersson, 1981). In this study, the method is extended to the situation where only some of the items are same between groups and each group has some unique items (referred to as *indicator shift* in Hancock et al., 2002). This is very common in the real testing situations. With this extension, multigroup factor analysis methods can then be applied in the areas where IRT methods dominate, such as equating and vertical scaling.

### 1.5 Purpose of the Study

The multidimensional nature of some multigroup assessments makes the application of unidimensional methods questionable. Although several multidimensional approaches have been proposed, the use of these methods has been limited in part because of the lack of knowledge with regard to which methods would be more appropriate under specific conditions and how these methods perform compared to the unidimensional methods. To date, little research has been conducted to explore these questions.

The purpose of this study was to compare and evaluate the performance of three multigroup multidimensional methods, specifically, the concurrent MIRT calibration, the separate MIRT calibration with linking, and the concurrent factor analysis calibration, under different conditions. The performance of unidimensional IRT models under these conditions was also investigated and compared with its multidimensional counterparts.

Note that only the models for binary (dichotomous) data were investigated in this study. The discussion of the models for polytomous data were beyond the scope of this study.



CHAPTER 2  
LITERATURE REVIEW  
MULTIGROUP IRT AND FACTOR ANALYSIS METHODS

In this chapter, the multigroup IRT and factor analysis methods are introduced. The relation between IRT and factor analysis methods is also discussed.

## 2.1. IRT Models

### 2.1.1 Unidimensional IRT (UIRT) Model and Linking

The unidimensional two-parameter logistic (2PL) IRT model (Lord & Novick, 1968) can be expressed as

$$P(X_i = 1|\theta) = \frac{e^{a_i(\theta-b_i)}}{1 + e^{a_i(\theta-b_i)}}, \quad (2-1)$$

where  $P(X_i = 1|\theta)$  is the probability of a correct response to item  $i$  given ability  $\theta$ ;  $a_i$  is the discrimination parameter for item  $i$ ; and  $b_i$  is the difficulty parameter for item  $i$ .

As has been discussed in Chapter 1, in the framework of IRT, the scale of parameters is determined only up to a linear transformation (Oshima et al., 2000). The probability of correct response is not altered by linear transformations (Hambleton et al., 1991)

$$a_i^* = \frac{a_i}{\alpha}, \quad (2-2)$$

$$b_i^* = \alpha b_i + \beta, \quad (2-3)$$

$$\theta_j^* = \alpha \theta_j + \beta, \quad (2-4)$$

where  $\alpha$  is a coefficient that adjusts the unit of the scale and  $\beta$  is a coefficient that adjusts the origin of the scale. It can be shown that

$$a_i^* (\theta_j^* - b_i^*) = \frac{a_i}{\alpha} ((\alpha\theta_j + \beta) - (\alpha b_i + \beta)) = \frac{a_i}{\alpha} (\alpha(\theta_j - b_i)) = a_i (\theta_j - b_i). \quad (2-5)$$

Therefore, when the IRT model holds, the scale of the parameter estimates from different groups are only linearly related (Kolen & Brennan, 2004). To put these estimates on the same scale, usually one group, for example group 1, is selected as the reference group and the scale of this group is used as the base scale, to which the parameters estimated in other groups are transformed,  $(\alpha, \beta)$  of the transformation equation are determined so that the parameter estimates from the other groups are as close as possible to the parameter estimates from group 1 after the transformation.

### 2.1.2 The Multidimensional IRT Model

Basically, there are two types of multidimensional IRT (MIRT) models: the *compensatory model* and the *noncompensatory model*. The compensatory models (Lord & Novick, 1968; McDonald, 1967; Reckase, 1985, 1995) allow the dimensions to interact: being low on one ability can be compensated for by being high on the other abilities to give a correct response. However, with the noncompensatory models (Embretson, 1984; Sympson, 1978), being low on one ability cannot be compensated for by being high on the other ability; one must demonstrate proficiency in all abilities in order to give a correct response. The current study focused on the more common compensatory models.

The multidimensional compensatory two-parameter logistic (MC2PL) model (Reckase, 1985) can be expressed as

$$P(X_i = 1|\boldsymbol{\theta}) = \frac{e^{(a_i'\boldsymbol{\theta} + d_i)}}{1 + e^{(a_i'\boldsymbol{\theta} + d_i)}}, \quad (2-6)$$

where  $P(X_i = 1|\boldsymbol{\theta})$  is the probability of a correct response to item  $i$  given ability  $\boldsymbol{\theta}$ ;  $\boldsymbol{\theta}$

is a  $n \times 1$  vector of ability parameters, where  $n$  is the number of dimensions;  $\mathbf{a}_i$  is a  $n \times 1$  vector of discrimination parameters;  $d_i$  is a scalar parameter that is related to the difficulty of the item. Note that

$$\mathbf{a}_i' \boldsymbol{\theta} + d_i = \sum_{k=1}^n a_{ik} (\theta_k - b_{ik}), \quad (2-7)$$

where  $a_{ik}$  is the  $k$  th element of  $\mathbf{a}_i$ , specifying the discrimination power of item  $i$  on dimension  $k$ ;  $\theta_k$  is the  $k$  th element of  $\boldsymbol{\theta}$ , specifying the ability on dimension  $k$ ;  $b_{ik}$  specifies the item difficulty on dimension  $k$ , and  $d_i = -\sum_{k=1}^n a_{ik} b_{ik}$ .

This model implies that the probability of a correct item response increases monotonically with the increase of the composite of the abilities on all dimensions. As the analog to the item characteristic curve (ICC) in the unidimensional IRT model, the relationship between the probability of correct response and the abilities can be graphically illustrated as an item characteristic surface (ICS). Figure 2-1 shows the ICS of a two dimensional MC2PL model, where  $a_1 = 1.0, a_2 = 0.5, d = 0.5$  (Bolt & Lall, 2003, with the change of some notations). As can be seen from the ICS, the probability is more sensitive to the change of  $\theta_1$ , which has a discrimination parameter of 1, than it is to the change of  $\theta_2$ , which has a discrimination parameter of .5.

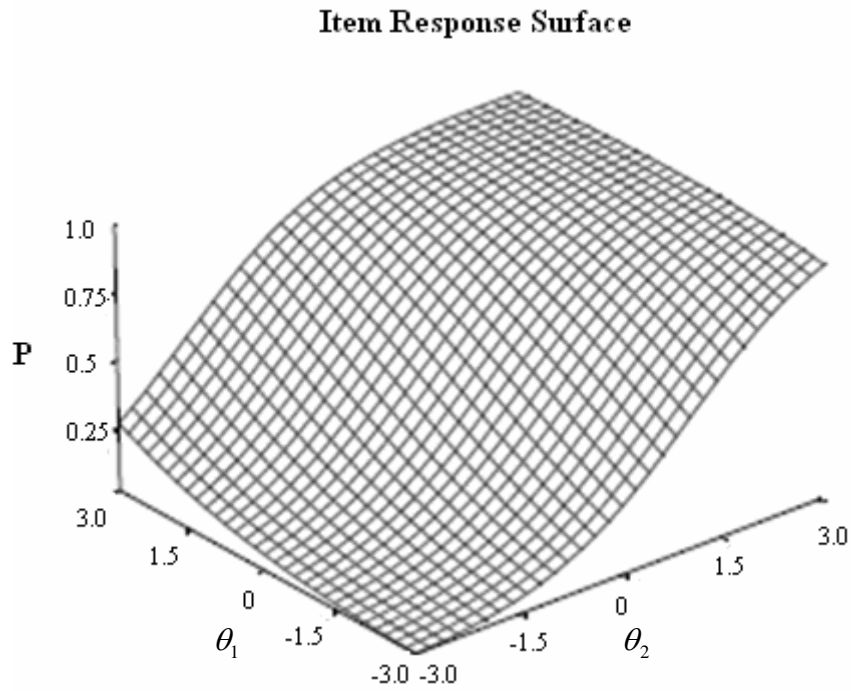


Figure 2-1 Item Response Surface (Resource: Bolt & Lall, 2003)

In MC2PL model, the analog to item discrimination and item difficulty in unidimensional IRT models is  $MDISC$  and  $MID$  (Reckase, 1985; Reckase & Mckinley, 1991). Graphically,  $MDISC$  represents the length of the discrimination vector in the ability space and can be calculated as

$$MDISC_i = \sqrt{\mathbf{a}'_i \mathbf{a}_i} = \left( \sum_{k=1}^n a_{ik}^2 \right)^{1/2}. \quad (2-8)$$

$MID$  represents the signed distance from the origin of the ability space to the point of the steepest slope on the ICS and can be calculated as

$$MID_i = \frac{-d_i}{MDISC_i}. \quad (2-9)$$

The direction of the item vector can be expressed as

$$\alpha_{ik} = \arccos\left(\frac{a_{ik}}{MDISC_i}\right), k = 1, 2, \dots, n, \quad (2-10)$$

where  $\alpha_{ik}$  is the angle of the item vector with dimension  $k$  for item  $i$ .

With the information of  $MDISC$ ,  $MID$ , and the direction of the item vector, the multidimensional items can be graphically displayed in the ability space. Note that if all item vectors were extended, they would pass through the origin. Figure 2-2 provides an example of two items in a two dimensional plane. In this example, item 1 is easier than item 2 and it has more discrimination power than item 2. Item 1 is more sensitive to  $\theta_2$  than to  $\theta_1$ . In contrast, item 2 is more sensitive to  $\theta_1$  than to  $\theta_2$ .

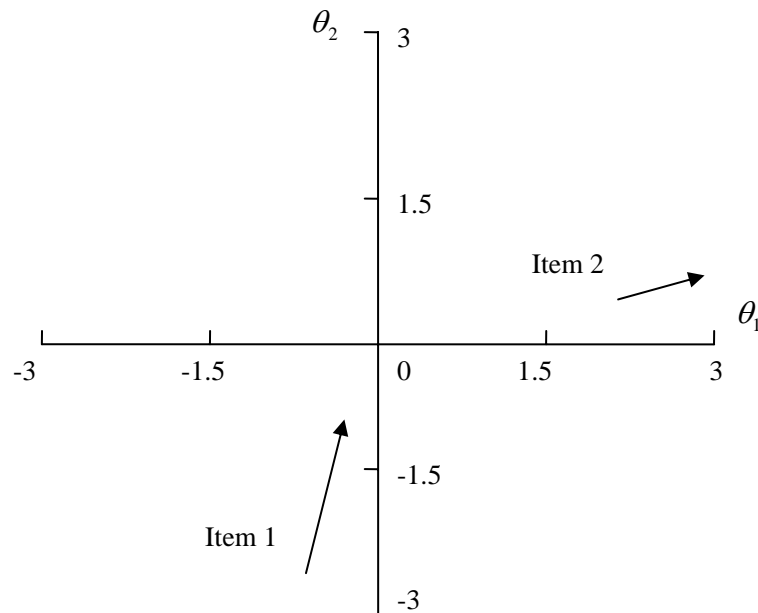


Figure 2-2 The item vectors in the ability space

### 2.1.3 Concurrent and separate MIRT calibration methods

#### BMIRT current and separate calibration

Yao (2003) developed a program, BMIRT, which can do both separate and concurrent parameter estimation for multigroup multidimensional IRT models.

BMIRT employs a Bayesian approach, which estimates the parameters based a

Markov chain Monte Carlo (MCMC) method (for more detail, see Yao & Daniel, 2006). BMIRT can do both exploratory and confirmatory analysis. In exploratory analysis, the number of dimensions can be determined by evaluating the change of model fit from each additional dimension. In the confirmatory analysis, the analysis is conducted based on the model that has been specified. In the output, BMIRT provides parameter estimates, model fit indices (e.g. chi-square, AIC, BIC), estimated score distribution for each group, and the estimated true score for each examinee.

In a simulation study conducted by Yao and Mao (2004) which compared the concurrent and separate calibration using BMIRT, it was found that the concurrent calibration always performed better than separate calibration.

#### NOHARM separate MIRT calibration

The Normal-Ogive Harmonic Analysis Robust Method (NOHARM) is a nonlinear item factor analysis method that can be used for single group multidimensional binary data analysis. The theory was developed by McDonald (1981, 1982, 1985) and programmed by Fraser and McDonald (1988). NOHARM approximates the MIRT normal-ogive model by a four-term polynomial series (for details see McDonald, 1983). Parameters are estimated using an unweighted least squares estimation based on the matrix of raw product moments. NOHARM can do both exploratory and confirmatory analysis. In exploratory analysis, an unrestricted model can be specified to obtain an exploratory solution, followed by either an orthogonal (Varimax) or oblique (Promax) rotation. The number of dimensions can be determined by evaluating the change of model fit from each additional dimension. In confirmatory analysis, the model can be described by specifying the parameters as either (1) fixed, (2) free to be estimated, or (3) constrained to be equal to one or

several other parameters. In all cases, NOHARM provides the parameter estimates as well as the matrix of covariance residuals. It also gives the root mean squares for the residual matrix as an overall measure of misfit of the model to the data. Note that NOHARM does not allow for missing data. So it is required that the data has been cleaned before running the analysis.

#### TESTFACT separate MIRT calibration

Full-information item factor analysis (Bock, Gibbons, & Schilling, 1988) provides another item factor analysis method for single group multidimensional data analysis. The method, implemented in TESTFACT (Bock et al., 1999), uses the marginal maximum likelihood (MML) estimation to provide full-information parameter estimates. That is, the estimates are based on all of the information in each examinee's pattern of correct and incorrect responses to all test items, not just the correct and incorrect frequencies for each item in the sample together with the joint correct and incorrect frequencies for all possible pairs of items (Toit, 2003). Details are given in Bock et al. (1988). TESTFACT can be used for exploratory factor analysis. The number of dimensions can be determined by evaluating the change of model fit from each additional dimension. TESTFACT also provides an option for confirmatory bifactor analysis (Holzinger & Swineford, 1937). The associated model assumes a single general dimension for all items plus one or more orthogonal "group" dimensions that also determine some or all of the items. Other than the confirmatory bifactor analysis, TESTFACT cannot be used for confirmatory factor analysis. The results provided by TESTFACT include a chi-square statistic for the model fit and the parameter estimates. In addition to full-information item factor analysis, TESTFACT can also do classical factor analysis based on tetrachoric correlations and uses the

estimates from a principal factor analysis of the tetrachoric correlation matrix as the starting value for full-information item factor analysis. Unlike NOHARM, TESTFACT allows missing data.

#### 2.1.4 Linking in MIRT

As has been discussed previously, scale indeterminacy also exists in MIRT models. Thus, the parameters estimated from different groups need to be transformed to a common scale, and the process is referred to as *multidimensional linking*.

##### The Davey, Oshima, and Lee Method

Davey (1991) introduced the theoretical background of a multidimensional linking method. For the multidimensional models with the exponent of  $\mathbf{a}'_i\boldsymbol{\theta} + d_i$ , the scale transformation can be conducted through the following transformation equations:

$$\mathbf{a}_i^* = (\mathbf{A}^{-1})' \mathbf{a}_i, \quad (2-11)$$

$$d_i^* = d_i - \mathbf{a}'_i \mathbf{A}^{-1} \boldsymbol{\beta}, \quad (2-12)$$

$$\boldsymbol{\theta}^* = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\beta}, \quad (2-13)$$

where  $\mathbf{A}$  is a  $n \times n$  rotation matrix ( $n$  is the number of dimensions), which has two functions: to rotate the orientation of the dimensions and to adjust the unit of the dimensions ; and  $\boldsymbol{\beta}$  is a  $n \times 1$  translation vector, which shifts the origin of a scale, So it can be shown that

$$\mathbf{a}_i^{*'} \boldsymbol{\theta}^* + d_i^* = \left( (\mathbf{A}^{-1})' \mathbf{a}_i \right)' (\mathbf{A}\boldsymbol{\theta} + \boldsymbol{\beta}) + (d_i - \mathbf{a}'_i \mathbf{A}^{-1} \boldsymbol{\beta}) = \mathbf{a}'_i \boldsymbol{\theta} + \mathbf{a}'_i \mathbf{A}^{-1} \boldsymbol{\beta} + d_i - \mathbf{a}'_i \mathbf{A}^{-1} \boldsymbol{\beta} = \mathbf{a}'_i \boldsymbol{\theta} + d_i. \quad (2-14)$$

Therefore, the transformation of the scale won't change the probability of



correct responses.

Davey and his colleagues (Davey, Oshima, & Lee, 1996; Oshima, Davey, Lee, 2000) proposed four procedures (the *direct method*, the *equated function method*, the *test characteristic function method*, and the *item characteristic function method*) to estimate  $(\mathbf{A}, \boldsymbol{\beta})$  in the transformation equation. These procedures, although employing slightly different criteria functions, are developed to make the corresponding parameter estimates from different scales as similar as possible after the transformation. All four methods estimate the rotation matrix and the translation vector simultaneously and allow non-orthogonal rotation of the matrix. A simulation study comparing the four methods (Oshima et al., 2000) suggested that linking from the test characteristic function (TCF) and the item characteristic function (ICF) methods are more stable than the other two procedures. In addition, the TCF method was best at estimating the rotation matrix over other three methods and was also relatively good at estimating the translation vector.

#### The Li and Lissitz Method

Li and Lissitz (2000) described another multidimensional linking method. In the Li and Lissitz's (2000) method, the dimensions are orthogonal. The scale linking consists of three parts: an orthogonal Procrustes rotation, a translation transformation, and a single dilation or contraction. The scale transformations are performed as follows

$$\mathbf{a}_i^* = k\mathbf{a}_i'\mathbf{T}, \quad (2-15)$$

$$d_i^* = d_i + \mathbf{a}_i'\mathbf{T}\mathbf{m}, \quad (2-16)$$

$$\boldsymbol{\theta}^* = (\mathbf{T}^{-1}\boldsymbol{\theta} - \mathbf{m})/k, \quad (2-17)$$

where  $\mathbf{T}$  is a  $n \times n$  orthogonal Procrustes rotation matrix, which rotates the orientation

of the dimensions;  $\mathbf{m}$  is a  $n \times 1$  translation vector, which shifts the origin of the scale; and  $k$  is a central dilation constant, which adjusts the unit of the dimensions. The equality of exponent terms after and before transformation is then established by

$$\mathbf{a}_i^* \boldsymbol{\theta}^* + d_i^* = (k\mathbf{a}_i' \mathbf{T})(1/k)(\mathbf{T}^{-1}\boldsymbol{\theta} - \mathbf{m}) + (d_i + \mathbf{a}_i' \mathbf{T}\mathbf{m}) = \mathbf{a}_i' \boldsymbol{\theta} - \mathbf{a}_i' \mathbf{T}\mathbf{m} + d_i + \mathbf{a}_i' \mathbf{T}\mathbf{m} = \mathbf{a}_i' \boldsymbol{\theta} + d_i \quad (2-18)$$

Whereas Davey et al.'s (1996, 2000) method estimate the transformation coefficients simultaneously, Li and Lissitz (2000) estimates the rotation matrix ( $\mathbf{T}$ ) and scaling coefficients ( $m$  and  $k$ ) separately.  $\mathbf{T}$  is estimated by minimizing the sum of squared differences between each pair of the corresponding item discrimination parameters from the two scales. Three sets of methods are proposed to estimate  $m$  and  $k$ . The two parameters can be simultaneously estimated by the *matching test response surfaces method*, or separately estimated with  $m$  by the *least squares for estimating translation parameters method*, and  $k$  by either the *ratio of eigenvalues for estimating the dilation parameter method* or the *ratio of trace for the dilation parameter method*. A simulation study (Li & Lissitz, 2000) indicated that Procrustes rotation satisfactorily estimated the rotation matrix; the least squares method produced a less biased and more stable estimate of  $m$  than the test response surfaces method; and the ratio of trace method performed best for the  $k$  estimation.

Min (2003) identified a limitation with Li and Lissitz's (2000) approach in that the scalar dilation parameter is insufficient for dilating the scales of the multiple dimensions. The scalar dilation adjusts the scale of different dimensions to exactly the same extent, but the separate calibration from different groups might dilate the scales of multiple dimensions to different degrees (Reckase & Martineau, 2003). To address this limitation, Min extended Li and Lissitz's (2000) transformation equations as:

$$\mathbf{a}_i^* = \mathbf{a}_i' \mathbf{T} \mathbf{K} \quad (2-19)$$

$$d_i^* = d_i + \mathbf{a}'_i \mathbf{T} \mathbf{m}, \quad (2-20)$$

$$\boldsymbol{\theta}^* = \mathbf{K}^{-1} (\mathbf{T}^{-1} \boldsymbol{\theta} - \mathbf{m}), \quad (2-21)$$

where  $\mathbf{K}$  is a diagonal dilation matrix and the elements on the diagonal of  $\mathbf{K}$  can be different, which allows for different dilation for different dimensions.

### The Reckase and Martineau Method

Reckase and Martineau (2004) identified an important weakness in the Min (2003) approach. When the number of dimensions is large, the computational load would be unfeasible. To solve this problem, Reckase and Martineau proposed employing an oblique Procrustes transformation method (Mulaik, 1972), which automatically aligns each dimension of the original matrix (comparison matrix) to the target matrix (base matrix) and, therefore, eliminates the need for a dilation parameter or vector. The rotation matrix from oblique Procrustes procedure is

$$\mathbf{T} = (\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}'\mathbf{B}, \quad (2-22)$$

where  $\mathbf{T}$  is the rotation matrix;  $\mathbf{A}$  is the comparison matrix; and  $\mathbf{B}$  is the base matrix.

The transformation equation is then

$$\mathbf{a}_i^* = \mathbf{a}'_i \mathbf{T}, \quad (2-23)$$

$$d_i^* = d_i + \mathbf{a}'_i \mathbf{T} \mathbf{m}, \quad (2-24)$$

$$\boldsymbol{\theta}^* = \mathbf{T}^{-1} \boldsymbol{\theta} - \mathbf{m}, \quad (2-25)$$

where  $\mathbf{m}$  is determined by minimizing sum of square difference between the estimate of  $d$  from the two groups after transformation.

As has been discussed earlier, Yon and Reckase (2005) compared Davey et al's (1996, 2000) method with Reckase and Martineau's (2004) method using both real and simulated data. Their study indicated that Reckase and Martineau's (2004)

method generally performed better than Davey et al.'s (1996, 2000) method in terms of parameter recovery.

## 2. 2 Factor analysis models

### 2.2.1 Factor analysis models for continuous data

The general factor analysis model with continuous indicator variables can be expressed as

$$\mathbf{Y} = \mathbf{\Lambda}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad (2-26)$$

where  $\mathbf{Y}$  is the  $p \times 1$  vector of observed indicator variables,  $p$  is the number of indicator variables;  $\mathbf{\Lambda}$  is a  $p \times n$  matrix of  $\lambda$  loadings,  $n$  is the number of factors (dimensions);  $\boldsymbol{\theta}$  is a  $n \times 1$  vector of factors; and  $\boldsymbol{\varepsilon}$  is a  $p \times 1$  vector of errors, which is assumed to follow  $N(\mathbf{0}, \boldsymbol{\Psi})$  when maximum likelihood estimation method is used;  $\boldsymbol{\Psi}$  is the  $p \times p$  diagonal matrix of the variance in  $\boldsymbol{\varepsilon}$ ;  $\boldsymbol{\theta}$  and  $\boldsymbol{\varepsilon}$  are independent with each other. The first and second order moment matrixes are then

$$E(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{\Lambda}E(\boldsymbol{\theta}), \quad (2-27)$$

$$E\left[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'\right] = \boldsymbol{\Sigma} = \mathbf{\Lambda}\boldsymbol{\Phi}\mathbf{\Lambda}' + \boldsymbol{\Psi}, \quad (2-28)$$

where  $\boldsymbol{\Phi}$  is the covariance matrix of  $\boldsymbol{\theta}$ .

The parameters can be estimated by comparing the observed and estimated first and second moment matrix through maximum likelihood estimation method (MLE), generalized least squares method (GLS), asymptotically distribution-free method (ADF), or other methods.

Factor analysis models also have the problem of indeterminacy.

Mathematically,

$$E(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{\Lambda}\mathbf{T}^{-1}E(\mathbf{T}\boldsymbol{\theta}) = \mathbf{\Lambda}\mathbf{T}^{-1}\mathbf{T}E(\boldsymbol{\theta}) = \mathbf{\Lambda}E(\boldsymbol{\theta}), \quad (2-29)$$

$$E\left[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'\right] = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\mathbf{T}^{-1}\mathbf{T}\boldsymbol{\Phi}\mathbf{T}'\mathbf{T}'^{-1}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}, \quad (2-30)$$

where  $\mathbf{T}$  is a matrix for linear transformation of the factor loading matrix  $\boldsymbol{\Lambda}$ .

In the single group analysis, the indeterminacy is usually removed by fixing one element as 1 and at least  $n - 1$  elements as 0 in each column of factor loading matrix  $\boldsymbol{\Lambda}$  (Jöreskog, 1971).

In multigroup analysis, when the measurement invariance assumption holds, the item parameters for the common items should be same across groups so that the first and second moment matrix of the common items of group  $k$  can be expressed as

$$E(\mathbf{Y}_{c(k)}) = \boldsymbol{\mu}_{c(k)} = \boldsymbol{\Lambda}_c E(\boldsymbol{\theta}_{c(k)}), \quad (2-31)$$

$$E\left[(\mathbf{Y}_{c(k)} - \boldsymbol{\mu}_{c(k)})(\mathbf{Y}_{c(k)} - \boldsymbol{\mu}_{c(k)})'\right] = \boldsymbol{\Sigma}_{c(k)} = \boldsymbol{\Lambda}_c \boldsymbol{\Phi}_{c(k)} \boldsymbol{\Lambda}_c' + \boldsymbol{\Psi}_{c(k)}, \quad (2-32)$$

where the subscript  $c$  indicates that the items are common items; and  $(k)$  indicates that the model is for group  $k$ . Note that there are no constraints on the unique items in each group during the process of estimation.

To remove the indeterminacy in multigroup analysis, usually one group is chosen as the reference group and the indeterminacy in the reference group is removed in the same way as in single group analysis. The indeterminacy in other groups is then removed by constraining the parameters of the common items to be equal to those in the reference group.

### 2.2.2 Factor analysis models for categorical data

In analyzing categorical data, factor analysis methods assume that there is a continuous latent response variable  $Y$  underlying each categorical variable  $X$  (Christofferson, 1975; Muthén, 1978; Muthén & Christofferson, 1981). The categorical item response is the result of categorizing the latent continuous variable

$Y$  by the threshold(s), as is illustrated in Figure 2-3. In this example, when the value of  $Y$  is greater than the threshold  $\tau$ , the observed dichotomous variable  $X$  is 1, otherwise it's 0.

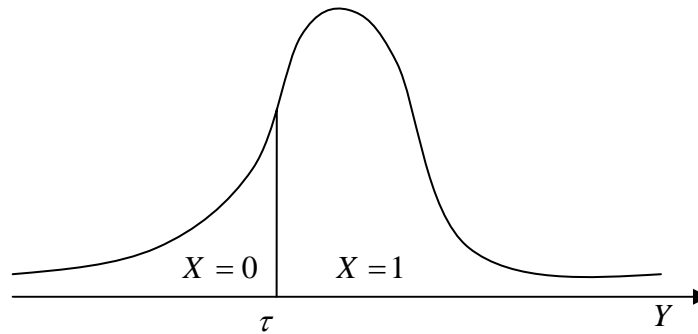


Figure 2-3 The categorization of the continuous  $Y$  into dichotomous  $X$

Therefore, the probability of  $X = 1$  is the probability of  $Y \geq \tau$ .

$$P(X = 1) = P(Y \geq \tau). \quad (2-33)$$

Knowing the relationship between  $X$  and  $Y$ , some important information about  $Y$  can be recovered by observing  $X$ . With the recovered information about  $Y$ , the relationship between  $Y$  and  $\theta$  can then be modeled by a regular factor analysis model for continuous data. Therefore, factor analysis for categorical data has two components: a threshold model describing the nonlinear relationship between  $X$  and  $Y$ , and an ordinary factor analysis model where  $Y$  is a linear function of  $\theta$  (Tate, 2003).

Let  $\mathbf{X}' = (X_1, X_2, \dots, X_p)$  be a random vector of responses to  $p$  dichotomous items. Assume that the joint distribution of  $\mathbf{Y}$  under these  $p$  items follows a multivariate normal distribution. Then the probability of observing response pattern  $\mathbf{X}$  is

$$P(\mathbf{X}) = \int_{R_1} \int_{R_2} \cdots \int_{R_p} f(\mathbf{Y}) d\mathbf{Y}, \quad (2-34)$$

where  $R_i$  of item  $i$  ( $i = 1, 2, \dots, p$ ) is the range of integral, which is  $[\tau_i, \infty)$  if  $X_i = 1$  and  $(-\infty, \tau_i)$  if  $X_i = 0$ ;  $\tau_i$  is the threshold for item  $i$ ;  $f(\mathbf{Y})$  is the joint density function and can be expressed as

$$f(\mathbf{Y}) = \frac{1}{|\boldsymbol{\Sigma}|^{p/2} (2\pi)^{p/2}} e^{-(\mathbf{Y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y}-\boldsymbol{\mu})/2}, \quad (2-35)$$

where  $\boldsymbol{\mu}$  is the mean vector of  $\mathbf{Y}$ ; and  $\boldsymbol{\Sigma}$  is the covariance matrix of  $\mathbf{Y}$ . Estimating the underlying continuous  $\mathbf{Y}$  also has a problem of scale indeterminacy, the unit and origin of the scale is quite arbitrary. Without loss of generality, in single group analysis, it is often assumed that  $\boldsymbol{\mu} = \mathbf{0}$  and  $\text{diag}(\boldsymbol{\Sigma}) = \mathbf{I}$  (Muthén & Christoffersson, 1981). The marginal distribution of  $Y_i$  of item  $i$  is

$$P_i = P(X_i = 1) = \int_{\tau_i}^{\infty} \frac{1}{(2\pi)^{1/2}} e^{-Y^2/2} dY, \quad (2-36)$$

$$Q_i = P(Y_i = 0) = 1 - P_i. \quad (2-37)$$

For a pair of items, item  $i$  and item  $j$ ,

$$P_{ij} = P(X_i = 1, X_j = 1) = \int_{\tau_i}^{\infty} \int_{\tau_j}^{\infty} \frac{1}{|\boldsymbol{\Sigma}_{ij}|^{1/2} 2\pi} e^{-\mathbf{Y}' \boldsymbol{\Sigma}_{ij}^{-1} \mathbf{Y}/2} d\mathbf{Y}. \quad (2-38)$$

Recall that the parameters in factor analysis models are estimated based on the first and second moment matrix of the indicator variables (the underlying  $\mathbf{Y}$  if the indicator variables are categorical). When  $\boldsymbol{\mu} = \mathbf{0}$  and  $\text{diag}(\boldsymbol{\Sigma}) = \mathbf{I}$ , the analysis are then based on the correlation matrix of  $\mathbf{Y}$ , which is referred to as the *tetrachoric correlation matrix* of  $\mathbf{X}$ . Estimating the tetrachoric correlation matrix and the thresholds of all items simultaneously from the observed item responses was computationally intensive at one time (Christoffersson, 1975). Christoffersson (1975) suggested estimating threshold of item  $i$  based on the marginal proportions  $P_i$ , and

tetrachoric correlation between item  $i$  and item  $j$  based on two-way joint proportion  $P_{ij}$ . The parameters can be estimated by maximum likelihood estimation (Bock & Lieberman, 1970) or generalized least squares estimation (Christoffersson, 1975; Muthén, 1978).

In multigroup analysis, usually one group is selected as the reference group, in which the mean and covariance matrix of  $\mathbf{Y}$  is constrained as  $\boldsymbol{\mu} = \mathbf{0}$  and  $diag(\boldsymbol{\Sigma}) = \mathbf{I}$ , the same as in single group analysis. The threshold  $\boldsymbol{\tau}$  and factor loadings  $\boldsymbol{\Lambda}$  of common items are constrained to be equal across groups as

$$\boldsymbol{\tau}^{(1)} = \boldsymbol{\tau}^{(2)} = \dots = \boldsymbol{\tau}^{(G)} = \boldsymbol{\tau}, \quad (2-39)$$

$$\boldsymbol{\Lambda}^{(1)} = \boldsymbol{\Lambda}^{(2)} = \dots = \boldsymbol{\Lambda}^{(G)} = \boldsymbol{\Lambda}. \quad (2-40)$$

As has been discussed in Chapter 1, most of the popular factor analysis programs (e.g., LISREL, Mplus, EQS) can do multigroup analysis for categorical data. In this study, Mplus was employed to conduct the analysis. Mplus estimates the parameters based on the maximum-likelihood estimation (for details, see Muthén & Asparouhov, 2002).

## 2.3. The equivalence of IRT and factor analysis method

### 2.3.1 Normal-ogive IRT model vs. logistic IRT model

Generally, there are two main variants of IRT models. One is the *normal-ogive IRT model* (“ogive” refers to the characteristic S-shape of the item response function). The other is the *logistic IRT model*. Although the normal-ogive model was dominant in early research on IRT, it has largely been replaced by the logistic model, which requires simpler computations (Crocker, 1986). The IRT models discussed in the previous sections are all logistic IRT models.



In the two-parameter normal-ogive binary IRT model (Bock & Aitkin, 1981; Bock & Lieberman, 1970), the probability of a correct response to item  $i$ , given ability  $\theta$ , is:

$$P(X_i = 1 | \theta) = \int_{-\infty}^{\infty} \phi(\mathbf{z}) d\mathbf{z} = \Phi(\mathbf{w}) \quad (2-41)$$

where  $\phi$  is the density function of the standard normal distribution;  $\mathbf{w} = \mathbf{a}'_i(\theta - \mathbf{b}_i)$ .

As in logistic IRT models,  $\mathbf{a}_i$  is the discrimination parameter (or vector of parameters in multidimensional models), and  $\mathbf{b}_i$  is the difficulty parameter ( $\mathbf{d}_i = \mathbf{a}'_i \mathbf{b}_i$  in multidimensional IRT model). Research (Haley, 1952; Lord & Novick, 1968) has proven that the relationship between the logistic distribution function  $L(\cdot)$  and the cumulative standard normal distribution  $F(\cdot)$  can be expressed as

$$|F(z) - L(1.7z)| < 0.01 \quad (2-42)$$

for all  $z$ . Therefore, the item parameters in the normal-ogive IRT model can be transformed to the corresponding parameters in the logistic IRT model by multiplying them by a scaling factor of 1.7. This is why logistic IRT model is often expressed as

$$P(X_i = 1 | \theta) = \frac{e^{1.7\mathbf{a}_i(\theta - \mathbf{b}_i)}}{1 + e^{1.7\mathbf{a}_i(\theta - \mathbf{b}_i)}} \quad (2-43)$$

In this case, the parameters  $\mathbf{a}_i$  and  $\mathbf{b}_i$  serve the same role in the logistic models as they do in the normal-ogive models.

### 2.3.2 The relationship between normal-ogive IRT model and factor analysis model

The relationship between normal-ogive IRT model and factor analysis model has been illustrated by Takane and Leeuw (1987) and Knol and Berger (1991). Recall that in the factor analysis model

$$\mathbf{Y} = \mathbf{\Lambda}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \quad (2-44)$$

Assume that  $\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{\varphi})$ , where  $\text{diag}(\boldsymbol{\varphi}) = \mathbf{I}$ ;  $\varepsilon_i \sim N(0, \psi_i)$ ; and  $\boldsymbol{\theta}$  and  $\varepsilon_i$  are independent of each other. It then follows that for item  $i$

$$Y_i \sim N(\mathbf{0}, \boldsymbol{\Lambda}'_i \boldsymbol{\varphi} \boldsymbol{\Lambda}_i + \psi_i). \quad (2-45)$$

The conditional distribution of  $Y_i$  given  $\boldsymbol{\theta}$  is

$$Y_i | \boldsymbol{\theta} \sim N(\boldsymbol{\Lambda}_i \boldsymbol{\theta}, \psi_i), \quad (2-46)$$

$$\frac{Y_i | \boldsymbol{\theta} - \boldsymbol{\Lambda}_i \boldsymbol{\theta}}{\sqrt{\psi_i}} \sim N(0,1). \quad (2-47)$$

The probability of a correct response to item  $i$  given  $\boldsymbol{\theta}$  is then

$$P(X_i = 1 | \boldsymbol{\theta}) = \int_{\tau_i}^{\infty} f(Y_i | \boldsymbol{\theta}) dY_i = \int_{\frac{\tau_i - \boldsymbol{\Lambda}_i \boldsymbol{\theta}}{\sqrt{\psi_i}}}^{\infty} f(z) dz = \Phi\left(\frac{\boldsymbol{\Lambda}_i \boldsymbol{\theta} - \tau_i}{\sqrt{\psi_i}}\right). \quad (2-48)$$

The normal-ogive IRT model is:

$$P(X_i = 1 | \boldsymbol{\theta}) = \Phi(\mathbf{a}'_i (\boldsymbol{\theta} - \mathbf{b}_i)). \quad (2-49)$$

Therefore,  $\mathbf{a}_i = \frac{\boldsymbol{\Lambda}_i}{\sqrt{\psi_i}}$  (2-50) and  $\mathbf{b}_i = \tau_i \boldsymbol{\Lambda}_i^{-1}$  (2-51). In multidimensional IRT model,

$d_i = -\frac{\tau_i}{\sqrt{\psi_i}}$  (2-52). Similar relationship can be found in Muthén (1979, Appendix),

Muthén & Christofferson (1981), and Bartholomew (1985).

## CHAPTER 3

### METHODOLOGY

Because the purpose of this study was not to explore the dimensionality structure of the data, confirmatory analysis approaches were investigated, which means that such information as the number of dimensions, the dimension(s) each item measured was already known before the analysis. The performance was evaluated based on the recovery of item parameters and the estimation of examinee true scores.

The study was based on simulated data because it is the best way to investigate the research questions in this study. In this chapter, statistical procedures for the simulation analysis are described. The criteria for evaluating the performance of the multigroup multidimensional methods are also provided.

#### 3.1 Simulation Design

In this study, a common item nonequivalent groups design was employed. In the design, it was assumed that two test forms, Form 1 and Form 2, were administered to two imaginary groups of examinees, Group 1 and Group 2. Each form consisted of 60 dichotomous items, 20 of which were common for both forms. Thus, there were 100 items in total for the two forms. Figure 3-1 illustrates the item composition of the two forms. The numbers in the parentheses are the number of items. Each form was developed to measure two abilities,  $\theta_1$  and  $\theta_2$ , so the latent structure of each form was two-dimensional. Assume also that the two abilities were compensatory with each other so that being low on one ability can be compensated for by being high on the other ability to give a correct response. In this study, guessing was assumed not to be a factor in getting a correct answer. Therefore, a two-dimensional MC2PL IRT model

was used to generate the response data. Remember that the MC2PL IRT model and factor analysis model are equivalent in the sense of formal mathematical functions. The response data generated from the IRT model were analyzed by both IRT methods and factor analysis methods. The parameters estimated from the factor analysis model were then translated to the IRT counterparts for comparison purposes.

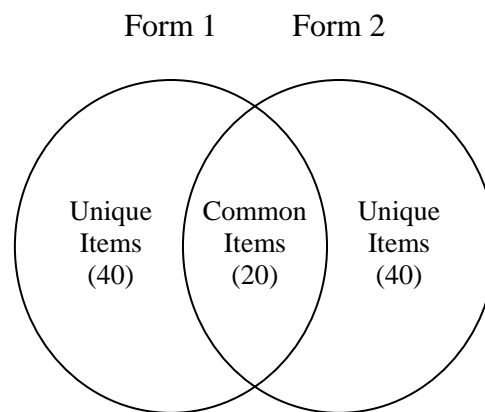


Figure 3-1 Items in Form 1 and Form 2

### 3.2 The Multigroup Analysis Methods Investigated

Four multigroup analysis methods were investigated. They were: concurrent MIRT calibration method, separate MIRT calibration method with linking, concurrent factor analysis calibration method, and concurrent UIRT calibration method. Both MIRT methods analyzed data based on MC2PL IRT model.

#### (1) Concurrent MIRT Calibration

The item parameters from the two forms were estimated simultaneously using BMIRT. The pooled data from both groups were analyzed in one step of analysis. In each run, the number of iterations was set to 5,000 and the burn-in was set to 2,000 (for each parameter, the estimate from the first 2,000 iterations was discarded, not used for the estimation of the distribution of the parameter).

## (2) Separate MIRT calibration with linking

The parameters were estimated separately for Form 1 and Form 2. NOHARM was employed for the calibration because TESTFACT can only be used for exploratory analysis or confirmatory analysis for the bifactor model (as discussed in Chapter 2), neither of which was the case in this study. Since Group 1 was selected as the reference group, the parameter scale in Group 1 was treated as the base scale, to which the parameters estimated in Group 2 were transformed.

## (3) Concurrent Factor Analysis Calibration

Factor analysis was carried out using Mplus. Group 1 was selected as the reference group with constraints for model identification. The parameters of common items were constrained to be equal across groups. The parameter estimates were transformed to the IRT scale through the transformation equations 2-50 and 2-52 discussed in Chapter 2.

## (4) Concurrent UIRT Calibration

The pooled data from Group 1 and Group 2 were analyzed in one run of BILOG-MG. The ability distribution of Group 1 was constrained to be a standard normal distribution. No constraints were imposed on Group 2.

As has been discussed in Chapter 2, the MIRT methods and factor analysis methods are equivalent when the dimensions follow a multivariate normal distribution. However, the two MIRT methods and the one factor analysis method employ different algorithms to estimate the parameters. Therefore, the differences, if any, between the performance from the different calibration methods are the result of using different estimation algorithms, not different models.

### 3.3 Key factors

This study examined how different factors affect the four calibration methods in multigroup analysis. Four factors were manipulated in the simulation. These factors were chosen based on the literature of simulation studies in related fields (Bolt, 1999, 2001; Dickenson, 2005; Finch, 2006; Kim, 2004; Oshima et al. ,2000; Spence, 1996; Tate, 2003).

#### (1) The Structural Orthogonality

The structural orthogonality was reflected by the correlation between the two dimensions ( $\theta_1$  and  $\theta_2$ ). Three levels of correlation were manipulated: 0.5, 0.7, and 0.9, which means  $\theta_1$  and  $\theta_2$  shared 25%, 49%, and 81% of variance.

#### (2) Equivalence of Test structure between Form 1 and Form 2

The equivalence of test structure was reflected by the equivalence of measurement emphasis on  $\theta_1$  and  $\theta_2$  in the two forms. The emphasis of the test was determined by the number of the items measuring  $\theta_1$  and  $\theta_2$  respectively.

Form 1 always had equivalent emphasis on  $\theta_1$  and  $\theta_2$ . In Form 1, 20 items measure  $\theta_1$  only, 20 items measure  $\theta_2$  only, and 20 items measure both. Among the 20 items measuring both  $\theta_1$  and  $\theta_2$ , 7 items were more sensitive to  $\theta_1$ , 7 were more sensitive to  $\theta_2$ , and 6 were equally sensitive to  $\theta_1$  and  $\theta_2$ . These 20 items were common items between the two forms. Figure 3-2 illustrates the orientation of the 60 items of Form 1 in the ability space. The degrees in parentheses represent the angle between the item vectors and  $\theta_1$ .

Three levels of measurement emphasis in Form 2 were manipulated as following:

a. Equivalent Emphasis

In this case, Form 2 had equal emphasis on  $\theta_1$  and  $\theta_2$  as Form 1 did, with 20 items measuring only  $\theta_1$ , 20 items measuring only  $\theta_2$ , and 20 items measuring both. The 20 items, measuring both  $\theta_1$  and  $\theta_2$ , were common items between Form 1 and Form 2. Those items measuring only  $\theta_1$  or  $\theta_2$  were unique items.

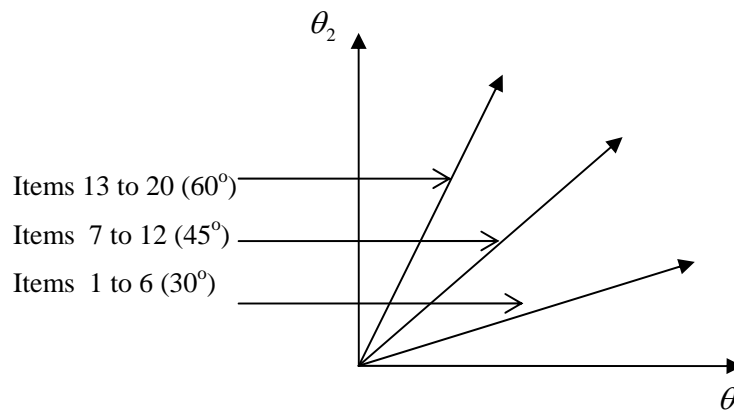


Figure 3-2 The distribution of the items of Form 1 in the ability space

b. Moderate Nonequivalent Emphasis

In this case, Form 2 had more emphasis on  $\theta_2$ , with 10 items measuring only  $\theta_1$ , 30 items measuring only  $\theta_2$ , and still 20 measuring both. Again, the 20 items measuring both  $\theta_1$  and  $\theta_2$  were common items.

c. Large Nonequivalent Emphasis

In this case, Form 2 put even more emphasis on  $\theta_2$  than condition b did. No item measured only  $\theta_1$ . In contrast, 40 items measured only  $\theta_2$ . The 20 items measuring both  $\theta_1$  and  $\theta_2$  in Form 2 were still common items, same as those in the above 2 scenarios.

Table 3-1 The number of items measuring  $\theta_1$  or  $\theta_2$  or both in Form 1 and Form 2 under different conditions

Nonequivalence in emphasis	Form 1			Form 2		
	$\theta_1$ only	Both	$\theta_2$ only	$\theta_1$ only	Both	$\theta_2$ only
No	20	20	20	20	20	20
Moderate	20	20	20	10	20	30
Large	20	20	20	0	20	40

The three conditions of structural equivalence are summarized in Table 3-1.

The first condition happens most frequently when two parallel forms of the same test are administered to different groups. The second and third conditions often happen when the two forms are from the tests of different grades and the emphasis of the tests changes according to the curriculum of the grades. For example, assume that grade 3 and grade 4 math tests both measure the abilities of data analysis and number sense. The grade 3 test has equal emphasis on both abilities, whereas the grade 4 test puts more emphasis on data analysis ability because the curriculum in grade 4 does so.

(3) Equivalence of item difficulty

Two levels of item difficulty equivalence for the two forms were manipulated.

a. Equivalent item difficulty

In this situation, the two forms had equivalent item difficulty. The mean and standard deviation of item difficulty parameter *MID* were 0 and 1 for both forms.

b. Nonequivalent item difficulty

In this situation, the unique items in Form 2 on average are more difficult than those in Form 1. The mean of the *MID* of the unique items in Form 2 is .5 higher than and those in Form 1. The standard deviation of the parameter did not change: it was still 1 for both forms.



(4) Equivalence of examinee groups

Three levels of the equivalence of the two groups of examinees were manipulated. It was assumed that the variance of  $\theta_1$  and  $\theta_2$  was 1 under all conditions. Thus, the two groups differed only in the mean proficiency on  $\theta_1$  and/or  $\theta_2$  when the two groups were not equivalent.

a. Equivalent on  $\theta_1$  and  $\theta_2$

In this case, the mean proficiency on  $\theta_1$  and  $\theta_2$  was 0 in both groups.

b. Not equivalent on  $\theta_2$

In this case, the mean proficiency on  $\theta_1$  was 0 in both groups, whereas the mean proficiency on  $\theta_2$  was 0 in Group 1 and was .5 in Group 2.

c. Not equivalent on both  $\theta_1$  and  $\theta_2$

In this case, Group 2 had higher mean proficiency on both  $\theta_1$  and  $\theta_2$ . The mean proficiency on both abilities were 0 in Group 1 and .5 in Group 2.

Table 3-2 The mean proficiency on  $\theta_1$  and  $\theta_2$  in the two groups

Group Equivalence	Group 1	Group 2
Equivalent on $\theta_1$ and $\theta_2$	( 0, 0)	( 0, 0)
Nonequivalent on $\theta_2$	( 0, 0)	( 0, .5)
Nonequivalent on $\theta_1$ and $\theta_2$	( 0, 0)	(.5, .5)

Table 3-2 summarizes the three conditions of group equivalence. A difference of 0.5 in the mean proficiency between the two groups was chosen because it is big enough to show the effect of difference (Li & Lissitz, 2000) and has been used in many simulation studies (Davey et al., 1996; Kim, 2004; Li & Lissitz, 2000; Min, 2003; Oshima et al. 2000; Skaggs & Lissitz, 1988).

When the two groups were equivalent, there was no need for linking in the separate calibration method because the standardization procedure (constraining the mean and standard deviation of  $\theta_1$  and  $\theta_2$  to be 0 and 1 within each group) had already put the parameter estimates on the same scale. In this case, the differences between the parameter estimates of the common items from the two forms were probably from sampling error. Therefore, the averages of the parameter estimates from the two forms were used as the parameter estimates for the common items in this study. When the two groups were not equivalent, however, linking was necessary for the separate calibration method because there were two sets of parameter estimates for the common items. One was from the estimation in Form 1. Another was from the estimation in Form 2 after transforming the parameter estimates to the scale of Form 1. In this case, the estimates from Group 1/Form 1 were used for the purpose of evaluation (Hanson & Beguin, 2002) since the scale of Form 1 was the base of the transformation, and the target of the transformation was to make the parameter estimates from Form 2 as close to those from Form 1 as possible. This was different from Kim and Cohen's (1998) method, where the average of the estimates of the common item parameters from the two groups were used to evaluate the parameter recovery.

In total, there were 54 combinations of conditions (3 structural orthogonality  $\times$  3 structural equivalence  $\times$  2 item difficulty equivalence  $\times$  3 examinee group equivalence). The conditions are summarized in Table 3-3. Under each condition, 100 replications were obtained in which all four methods converged. The value of 100 was chosen because it is common in simulation studies in related fields (Dickenson, 2005; Li & Lissitz, 2000; Kim, 2004). In some cases, Mplus failed to converge. When this

happened, a new set of data were generated and the four methods applied. This process was continued until 100 successes were obtained.

### 3.4 Data generation

#### 3.4.1 Item parameters generation

Remember that the multidimensional compensatory two-parameter logistic (MC2PL) model (Reckase, 1985) can be expressed as

$$P(X_i = 1|\boldsymbol{\theta}) = \frac{e^{(\mathbf{a}'_i\boldsymbol{\theta}+d_i)}}{1 + e^{(\mathbf{a}'_i\boldsymbol{\theta}+d_i)}}, \quad (3-1)$$

where  $\boldsymbol{\theta}$  is a vector of ability parameters;  $\mathbf{a}_i$  is a vector of discrimination parameters; and  $d_i$  is a scalar parameter that is related to the difficulty of the item.

$MDISC$  and  $MID$  are two parameters derived from MC2PL model. They represent the overall item discrimination and item difficulty for the item,

$$MDISC_i = \sqrt{\mathbf{a}'_i\mathbf{a}_i} = \left( \sum_{k=1}^n a_{ik}^2 \right)^{1/2}, \quad (3-2)$$

$$MID_i = \frac{-d_i}{MDISC_i}. \quad (3-3)$$

In this study  $MDISC$  and  $MID$  were generated first. The parameters in MC2PL model were then determined based on  $MDISC$ ,  $MID$ , and  $\alpha_{i1}$  (the angle between the item vector and  $\theta_1$ ).

#### (1) Item Discrimination Parameter $MDISC$

The literature from previous simulation studies indicated that the researchers had different beliefs about the distribution of  $MDISC$ . Although most of the researchers believed that  $MDISC$  follows the lognormal distribution (Bolt, 2001;

Dickenson, 2005; Finch, 2006; Min, 2003; Skaggs & Lissitz, 1988; Spence, 1996; Tate, 2003; Yao, 2006), there is still a lot of disagreement about the reasonable value range of the parameter. In this study, *MDISC* was assumed to follow a lognormal distribution. The mean and standard deviation of  $\log(MDISC)$  distribution were set to 0 and .5, which are the default values of the distribution of  $a$  in BILOG-MG program. Because there were 100 items in total, 100 *MDISC* values were randomly generated from this lognormal distribution with the range of .5 to 2.5, which was chosen according to the results of empirical studies reported by Doody-Bogan and Yen (1983), Ackerman (1988), Spence (1996), and Roussos et al. (1998). Of the 100 *MDISC* value generated, 20 were randomly selected for the common items, 40 for the unique items of Form 1, and the other 40 for the unique items of Form 2.

Because the 20 common items measured both  $\theta_1$  and  $\theta_2$ , the value of  $a_{i1}$  and  $a_{i2}$  were determined by  $\alpha_{i1}$  and *MDISC*

$$a_{i1} = MDISC \times \cos(\alpha_{i1}), \quad (3-3)$$

$$a_{i2} = MDISC \times \sin(\alpha_{i1}). \quad (3-4)$$

Note that of the 20 common items, the  $\alpha_{i1}$  of 7 items were  $30^\circ$ , 6 items were  $45^\circ$ , and the other 7 items were  $60^\circ$ .

The 40 unique items of each form had simple structure. In Form 1, 20 items measured only  $\theta_1$  so the discrimination vectors for these items were in the form of  $(a_{i1}, 0)$ , with  $a_{i1} = MDISC_i$ . The other 20 items measured only  $\theta_2$  so the discrimination vectors for these items were in the form of  $(0, a_{i2})$ , with  $a_{i2} = MDISC_i$ . Three versions of Form 2 were generated according to the change of the measurement emphasis of the test. In the first version, Form 2 had equivalent emphasis on  $\theta_1$  and  $\theta_2$ , same as Form 1, with discrimination vectors of 20 unique items as  $(a_{i1}, 0)$  and the other 20

items as  $(0, a_{i2})$ . In the second version, Form 2 had more emphasis on  $\theta_2$ . The discrimination vector of 10 items, randomly selected from the 20 items measuring  $\theta_1$  in the first version, were changed from  $(a_{i1}, 0)$  to  $(0, a_{i2})$ , with  $MDISC_i$  unchanged. In the third version, Form 2 put even more emphasis on  $\theta_2$  than the second version; all unique items of Form B measured only  $\theta_2$ , with discriminatio vectors as  $(0, a_{i2})$ .

## (2) Item Difficulty Parameter $MID$

Item difficulty parameter  $MID$  was assumed to follow the normal distribution by many researchers (Bolt, 2001; Finch, 2006; Spence, 1996; Yao, 2006). The range of  $MID$  in this study was determined based on the previous studies so that it's reasonable for published tests. 100  $MID$  values were randomly generated from a standard normal distribution with the range from -2 to 2 (Finch, 2006; Spence, 1996). These values were randomly assigned to the 100 items (20 common items, 40 unique items of Form 1, and 40 unique items of Form 2). This was the case for equivalent item difficulty for the two forms. When Form 2 was more difficult than Form 1, each  $MID$  value originally generated for the unique items in Form 2 was increased by .5 so that the average difficulty level of Form 2 is higher than Form 1. The value of  $d_i$  in the MC2PL model was calculated by

$$d_i = -MDISC_i \times MID_i . \quad (3-5)$$

### 3.4.2. Generation of Correlated $\theta_1$ and $\theta_2$ for Group 1 and Group 2

A sample size of 2,000 or more is usually suggested for MIRT calibration (Akerman, 1994; Reckase, 1995), which indicates that MIRT methods are more suitable for large scale assessment. In this study, the sample size for Group 1 and

Group 2 were both 2,000.

To generate correlated  $\theta_1$  and  $\theta_2$  values for each examinee, the following procedure was followed. Assuming the values of  $\theta_1$  and  $\theta_2$  were determined by a higher-order standard normal random variable  $z$ , as is shown in Figure 3-3.

$$\begin{aligned}\theta_1 &= \alpha_1 + \beta z + \sqrt{1 - \beta^2} \zeta_1 \\ \theta_2 &= \alpha_2 + \beta z + \sqrt{1 - \beta^2} \zeta_2\end{aligned}, \quad (3-6)$$

where  $\alpha_1$  and  $\alpha_2$  were the mean of  $\theta_1$  and  $\theta_2$  in the population;  $\beta^2$  equals the targeted correlation between  $\theta_1$  and  $\theta_2$ ; and  $\zeta_1$  and  $\zeta_2$  are two standard normal random variables.

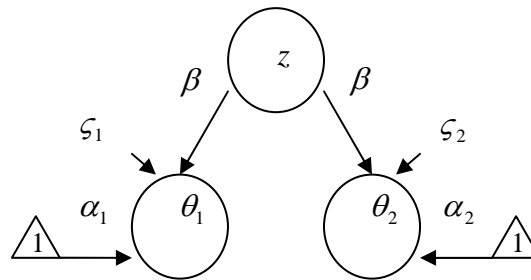


Figure 3-3 Generating correlated  $\theta_1$  and  $\theta_2$  from higher order variable  $z$

By this means, 2,000 pairs of correlated  $\theta_1$  and  $\theta_2$  were generated for each group in every replication of the simulation. Note that the value of  $\alpha_1$ ,  $\alpha_2$ , and  $\beta$  were determined by the value of manipulated factors.

### 3.4.3 Generate Response Data

With the generated item and person parameters, the probability of correct response to item  $i$  by person  $j$ ,  $P_{ij}$ , were calculated from the MC2PL model.

To generate the 0 / 1 response,  $X_{ij}$ , a uniform random number,  $R$ , was

generated in the range of (0, 1). Comparing  $R$  with  $P_{ij}$ , the value of  $X_{ij}$  was determined by the following rule:

$$X_{ij} = \begin{cases} 0, & R > P_{ij} \\ 1, & R \leq P_{ij} \end{cases} . \quad (3-7)$$

In this study, all parameter values and response data were generated by the SAS program.

### 3.5 Linking for Separate MIRT Calibration

In Chapter 2, several methods of linking for separate MIRT calibration have been discussed. Davey et al.'s (1996, 2000) methods allow non-orthogonal rotation of the dimensions. The rotation matrix takes care of the orientation and the unit of the dimensions simultaneously. Li and Lissitz's (2000) method assumes that the dimensions are orthogonal. The orientation of the dimensions is rotated by an orthogonal Procrustes rotation matrix. The unit of the dimensions is adjusted by a central dilation constant. Min's (2003) method extends Li and Lissitz's (2000) method by replacing the dilation constant with a diagonal dilation matrix. This change allows the units of the different dimensions to be adjusted to different levels. Reckase and Martineau's (2004) method employs an oblique Procrustes transformation (Mulaik, 1972) approach, which automatically aligns each dimension of the original matrix (comparison matrix) to the target matrix (base matrix) and, therefore, eliminates the need for a dilation parameter or vector.

Confirmatory analysis was employed in this study, which means that the information of the dimension(s) measured by each item was already known at the beginning of the analysis. In Form 1 and Form 2, the unique items measure only one of the two dimensions so that the discrimination vectors of these items were in the

form of  $(a_1, 0)$  or  $(0, a_2)$ . In this way, the direction of the dimensions was determined and there was no need for rotation. The only indeterminacies remaining were then the unit and origin of the dimensions. In this study, Min's (2003) method was employed for linking with some changes. This method was chosen because it takes care of the dimension orientation and unit separately so that the unit of the dimension could be adjusted without changing the orientation of the dimensions. In addition, the method allows the dilation to be different across dimensions. The scale transformation employed in this study was then

$$\mathbf{a}_i^* = \mathbf{a}_i' \mathbf{E} \mathbf{K}, \quad (3-8)$$

$$d_i^* = d_i + \mathbf{a}_i' \mathbf{E} \mathbf{m}, \quad (3-9)$$

$$\boldsymbol{\theta}^* = \mathbf{K}^{-1}(\mathbf{E}\boldsymbol{\theta} - \mathbf{m}). \quad (3-10)$$

Note that the orthogonal Procrustes rotation matrix  $\mathbf{T}$  in Min's (2003) method was replaced by an identity matrix  $\mathbf{E}$  so that no change was made to the orientation of the dimensions. The estimation of  $\mathbf{K}$  and  $\mathbf{m}$  is described in detail in the Appendix A.

### 3.6 Evaluation Criteria

The performance of the methods was evaluated from two perspectives. The first one evaluated the recovery of item parameters by comparing the estimates from different calibration methods with the true value of the parameters. This is the most often used criterion in the simulation studies conducted previously to investigate the performance of calibration methods (Kim, 2004; Li & Lissitz, 2000; Min, 2003; Oshima et al., 2000). Note that this criterion was not applicable to the unidimensional model because there were no true unidimensional parameters to compare with. The second criterion was about the accuracy of the estimation of true score (the model-indicated total score) of examinees. The performance was evaluated based on the



difference between the estimated true scores from the calibration methods and the “true” true scores obtained with the model used to generate the data. This criterion was not applicable to the separate MIRT calibration in this study because NOHARM does not provide the estimation of  $\theta_1$  and  $\theta_2$ . Therefore no estimated true score was available.

(1) Recovery of Item Parameters

a. *BIAS*

*BIAS* is a measure of the accuracy of the estimation of parameter. It was calculated by taking the mean differences between the true parameter values and the corresponding estimates over the 100 iterations. For a given parameter, for example  $d_i$  of item  $i$ ,  $BIAS_i$  can be calculated as

$$BIAS_i = \frac{1}{100} \sum_{n=1}^{100} (\hat{d}_{in} - d_{i,true}), \quad (3-11)$$

where  $\hat{d}_{in}$  is the estimate of  $d_i$  in the  $n$ th replication;  $d_{i,true}$  is the true value of  $d_i$ .

b. Standard deviation (*SD*) of parameter estimate

*SD* is a measure of the stability of estimation. Again, use  $d_i$  as an example,  $SD_i$  can be calculated as

$$SD_i = \sqrt{\frac{1}{99} \sum_{n=1}^{100} (\hat{d}_{in} - \bar{\hat{d}}_i)^2}, \quad (3-12)$$

where  $\bar{\hat{d}}_i = \frac{1}{100} \sum_{n=1}^{100} \hat{d}_{in}$ .

## (2) Estimation of True Score of Examinees

The “true” true score for person  $j$  in the  $n$  th replication can be estimated as

$$T_{nj} = \sum_{i=1}^{60} P(X_{nij} = 1). \quad (3-13)$$

### a. *BIAS*

*BIAS* measures the average of the difference between the estimated true score and “true” true score for all examinees in the group. The *BIAS* of the true score estimation in one group over 100 replication can be calculated as

$$BIAS = \frac{1}{100 \times 2000} \sum_{n=1}^{100} \sum_{j=1}^{2000} (\hat{T}_{nj} - T_{nj}), \quad (3-14)$$

where  $\hat{T}_{nj}$  is the estimated true score of person  $j$  given the estimated parameters.

### b. *SD*

*SD* reflects the variability of the difference between the estimated and “true” true score among the examinees in the group. *SD* of the true score estimation in one group over 100 replication can be calculated as

$$SD = \sqrt{\frac{1}{100 \times 2000 - 1} \sum_{n=1}^{100} \sum_{j=1}^{2000} (\hat{T}_{nj} - T_{nj})^2}. \quad (3-15)$$

## CHAPTER 4

### RESULTS

This chapter presents the results from the simulation study described in Chapter 3. The performance of the three multidimensional calibration methods (concurrent MIRT, separate MIRT with linking, concurrent Factor Analysis) and their unidimensional counterpart (concurrent UIRT) were investigated under different conditions. The performance was evaluated based on the recovery of the item parameters and the estimation of the true score of examinees. The effect of the four manipulated factors (the structural orthogonality, the equivalence of test structure, item difficulty, and examinee groups) on the performance of the four methods was also investigated.

Table 4-1 summarizes the 54 combinations of conditions from the four manipulated factors in this study. In the table, the “correlation” column represents the three levels of the structural orthogonality, reflected by the correlation between  $\theta_1$  and  $\theta_2$ , which were 0.5, 0.7, and 0.9 respectively. The “emphasis” column represents the three levels of structural equivalence between Form 1 and Form 2, reflected by the measurement emphasis on  $\theta_1$  and  $\theta_2$  in the forms. The three numbers in the parenthesis are the number of items measuring both  $\theta_1$  and  $\theta_2$ , only  $\theta_1$ , and only  $\theta_2$  in Form 2. For example, (20, 10, 30) means that of the 60 items in Form 2, 20 items measure both  $\theta_1$  and  $\theta_2$ , 10 items measure only  $\theta_1$ , and the other 30 items measure only  $\theta_2$ . Note that the measurement emphasis on  $\theta_1$  and  $\theta_2$  was always equivalent in Form 1 so that the items were (20, 20, 20) under all conditions. The “difficulty” column represents the two levels of item difficulty equivalence between the two forms. “Equivalent” means that the two forms had equivalent item difficulty. “.5 higher”

means that the mean of *MID* of the unique items in Form 2 was .5 higher than that in Form 1. The “ability” column represents the three levels of equivalence of the two examinee groups. The mean of  $\theta_1$  and  $\theta_2$  in Group 1 was (0, 0) for all conditions. Three levels of the mean of  $\theta_1$  and  $\theta_2$  in Group 2 were (0, 0), (0, .5), and (.5, .5). Same notations are used in all figures and tables in this chapter and the appendix.

Table 4-1 The 54 combinations of conditions

Condition	Correlation	Emphasis	Difficulty	Ability
1	0.5	(20,20,20)	equivalent	(0,0)
2	0.5	(20,20,20)	equivalent	(0,.5)
3	0.5	(20,20,20)	equivalent	(.5,.5)
4	0.7	(20,20,20)	equivalent	(0,0)
5	0.7	(20,20,20)	equivalent	(0,.5)
6	0.7	(20,20,20)	equivalent	(.5,.5)
7	0.9	(20,20,20)	equivalent	(0,0)
8	0.9	(20,20,20)	equivalent	(0,.5)
9	0.9	(20,20,20)	equivalent	(.5,.5)
10	0.5	(20,20,20)	.5 higher	(0,0)
11	0.5	(20,20,20)	.5 higher	(0,.5)
12	0.5	(20,20,20)	.5 higher	(.5,.5)
13	0.7	(20,20,20)	.5 higher	(0,0)
14	0.7	(20,20,20)	.5 higher	(0,.5)
15	0.7	(20,20,20)	.5 higher	(.5,.5)
16	0.9	(20,20,20)	.5 higher	(0,0)
17	0.9	(20,20,20)	.5 higher	(0,.5)
18	0.9	(20,20,20)	.5 higher	(.5,.5)
19	0.5	(20,10,30)	equivalent	(0,0)
20	0.5	(20,10,30)	equivalent	(0,.5)
21	0.5	(20,10,30)	equivalent	(.5,.5)
22	0.7	(20,10,30)	equivalent	(0,0)
23	0.7	(20,10,30)	equivalent	(0,.5)
24	0.7	(20,10,30)	equivalent	(.5,.5)
25	0.9	(20,10,30)	equivalent	(0,0)
26	0.9	(20,10,30)	equivalent	(0,.5)
27	0.9	(20,10,30)	equivalent	(.5,.5)

Table 4-1 The 54 combinations of conditions (continued)

Condition	Correlation	Emphasis	Difficulty	Ability
28	0.5	(20,10,30)	.5 higher	(0,0)
29	0.5	(20,10,30)	.5 higher	(0,.5)
30	0.5	(20,10,30)	.5 higher	(.5,.5)
31	0.7	(20,10,30)	.5 higher	(0,0)
32	0.7	(20,10,30)	.5 higher	(0,.5)
33	0.7	(20,10,30)	.5 higher	(.5,.5)
34	0.9	(20,10,30)	.5 higher	(0,0)
35	0.9	(20,10,30)	.5 higher	(0,.5)
36	0.9	(20,10,30)	.5 higher	(.5,.5)
37	0.5	(20, 0,40)	equivalent	(0,0)
38	0.5	(20, 0,40)	equivalent	(0,.5)
39	0.5	(20, 0,40)	equivalent	(.5,.5)
40	0.7	(20, 0,40)	equivalent	(0,0)
41	0.7	(20, 0,40)	equivalent	(0,.5)
42	0.7	(20, 0,40)	equivalent	(.5,.5)
43	0.9	(20, 0,40)	equivalent	(0,0)
44	0.9	(20, 0,40)	equivalent	(0,.5)
45	0.9	(20, 0,40)	equivalent	(.5,.5)
46	0.5	(20, 0,40)	.5 higher	(0,0)
47	0.5	(20, 0,40)	.5 higher	(0,.5)
48	0.5	(20, 0,40)	.5 higher	(.5,.5)
49	0.7	(20, 0,40)	.5 higher	(0,0)
50	0.7	(20, 0,40)	.5 higher	(0,.5)
51	0.7	(20, 0,40)	.5 higher	(.5,.5)
52	0.9	(20, 0,40)	.5 higher	(0,0)
53	0.9	(20, 0,40)	.5 higher	(0,.5)
54	0.9	(20, 0,40)	.5 higher	(.5,.5)

#### 4.1 Recovery of the item parameters

The evaluation of the recovery of the item parameters was only available for the three multidimensional calibration methods. Two criteria were used: *BIAS* measured the average difference between the estimated and the true value of the parameters; *SD* reflected the stability of the estimation. The detailed information about the *BIAS* and *SD* of  $a_1$ ,  $a_2$ , and  $d$  can be found in the tables in Appendix B. In the tables are the averages of the *BIAS* or *SD* over the 100 items under each condition. The bold-faced numbers in the tables indicate the methods that resulted in

the smallest *BIAS* or *SD* under each condition.

#### 4.1.1 The recovery of $a_1$

Figures 4-1 and Figure 4-2 depict the *BIAS* and the *SD* of  $a_1$  from the three methods under all 54 conditions. In the figures, the solid line represents the concurrent MIRT calibration, which was carried out in the program BMIRT; the dashed line represents the separate MIRT calibration with linking, which was carried out in the program NHOARM; the dotted line represents the concurrent factor analysis calibration, which was carried out in the program Mplus. For the ease of illustration, in this chapter, the calibration methods were represented by the name of the computer programs that carried out the analysis. Specifically, BMIRT represents the concurrent MIRT calibration; NOHARM represents the separate MIRT calibration with linking; Mplus represents the concurrent factor analysis calibration. Each figure is split into three parts based on the three levels of the equivalence of test structure between the two forms. In addition, the conditions with relatively large *BIAS* are labeled with the corresponding value of the factor(s) that is(are) common to these conditions.

Figures 4-3 to 4-6 depict the effect of the four manipulated factors on the *BIAS* and the *SD* of the estimates of  $a_1$  from the three methods. Boxplots were employed. Each bounded vertical line represents the range of the observations in a set of data for each method at each factor level. The bottom and the top of the box represent the first quartile (Q1) and the third (Q3) quartile of the data. The horizontal line in the middle of the box represents the median. The dots away from the box, with the condition numbers, are outliers. The abscissa of each plot represents the levels of the specific factor. The cluster of three boxes represents the observations from the

three methods under each level of the factor.

Therefore, Figures 4-1 and 4-2 reflect the main effect of the calibration methods on the estimate of  $a_1$  and some higher-order interaction effects between the calibration methods and the manipulated factors. Figures 4-3 to 4-6 reflect the first order interaction effects between the calibration methods and the manipulated factors.

From Figures 4-1 to 4-6, it can be found that the three methods performed differently on the estimation of  $a_1$ . The following are the major findings.

- (1) In general, the *BIAS* of  $a_1$  from Mplus and NOHARM were comparable and close to zero under most conditions, which indicates that the estimate of  $a_1$  from the two methods, on average, were very close to their true value. BMIRT tended to underestimate  $a_1$  under all conditions and the absolute magnitude of *BIAS* was larger than that from the other two methods.
- (2) The *SD* of  $a_1$  from NOHARM was generally larger than that from the other two methods and tended to fluctuate widely across conditions. The *SD* from Mplus also fluctuated across conditions, but with a smaller magnitude than that from NOHARM. The *SD* of  $a_1$  from BMIRT was the smallest under most conditions, and it tended to be more consistent across conditions, which indicates that it was less affected by the manipulated factors.
- (3) When the correlation between  $\theta_1$  and  $\theta_2$  increased, the absolute magnitude of the *BIAS* from BMIRT increased. For all three methods, the estimate of  $a_1$  became less stable as the correlation increased, especially when the correlation increased from 0.7 to 0.9.

- (4) When the “emphasis” in Form 2 was (20, 10, 30), the estimate of  $a_1$  from Mplus tended to be more stable than when the emphasis was (20, 20, 20) or (20, 0, 40).
- (5) When the two groups were not equivalent, specifically, when the “ability” was (0, .5) or (.5, .5), the estimate of  $a_1$  from NOHARM became less stable.
- (6) There were some higher order interaction effects among the manipulated factors and the calibration methods. When the “emphasis” was (20, 0, 40) and the “Ability” was (0, 0), NOHARM tended to underestimate the parameter. When the “emphasis” was (20, 20, 20) or (20, 10, 30), the *SD* from Mplus was relatively large when the “correlation” was 0.9 and the ability was (0, .5) or (.5, .5). When the “emphasis” was (20, 0, 40), the *SD* from Mplus was relatively large when the “correlation” was 0.9.



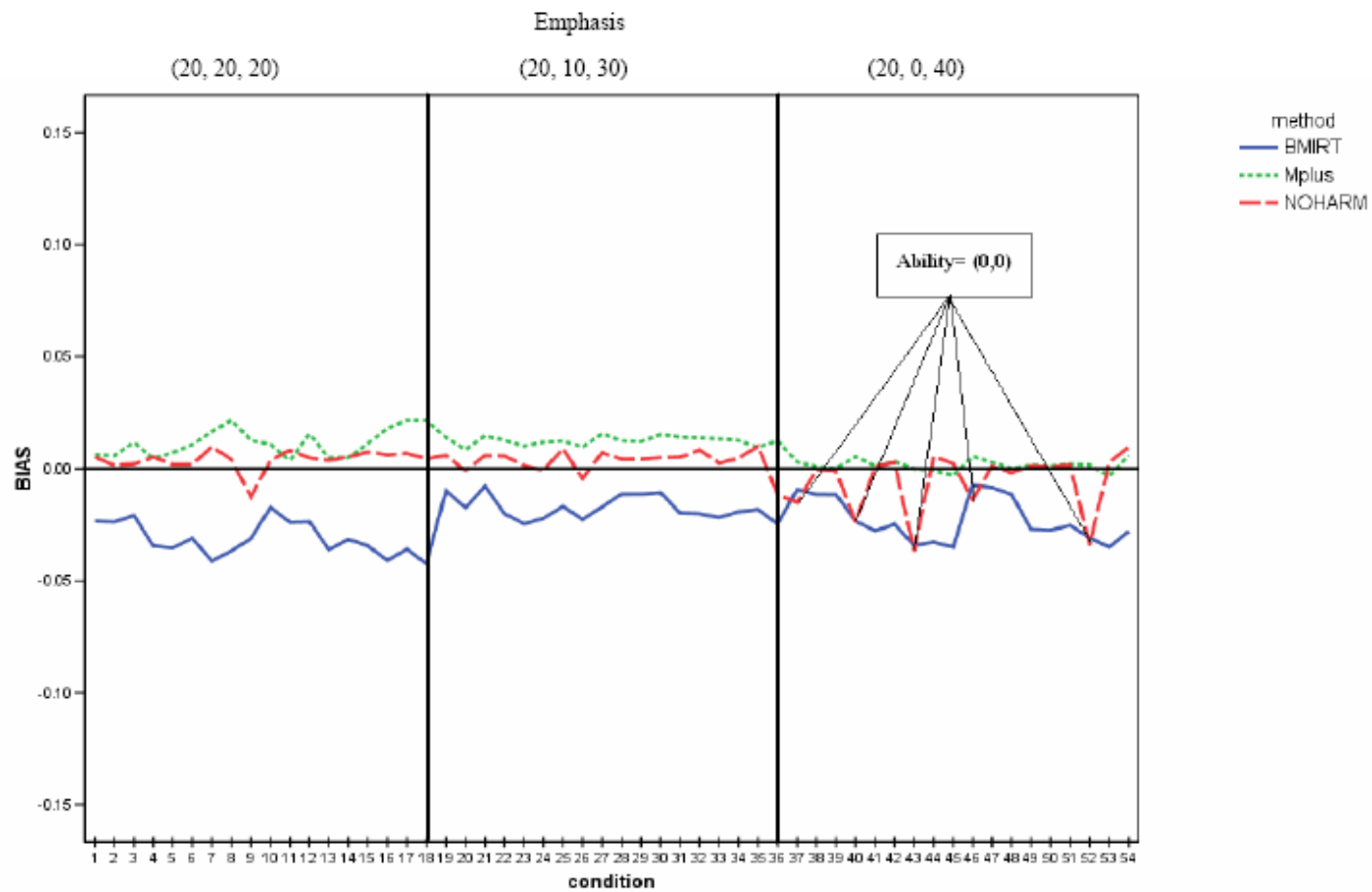


Figure 4-1 *BIAS* of  $a_1$  from the three calibration methods  
under all 54 conditions

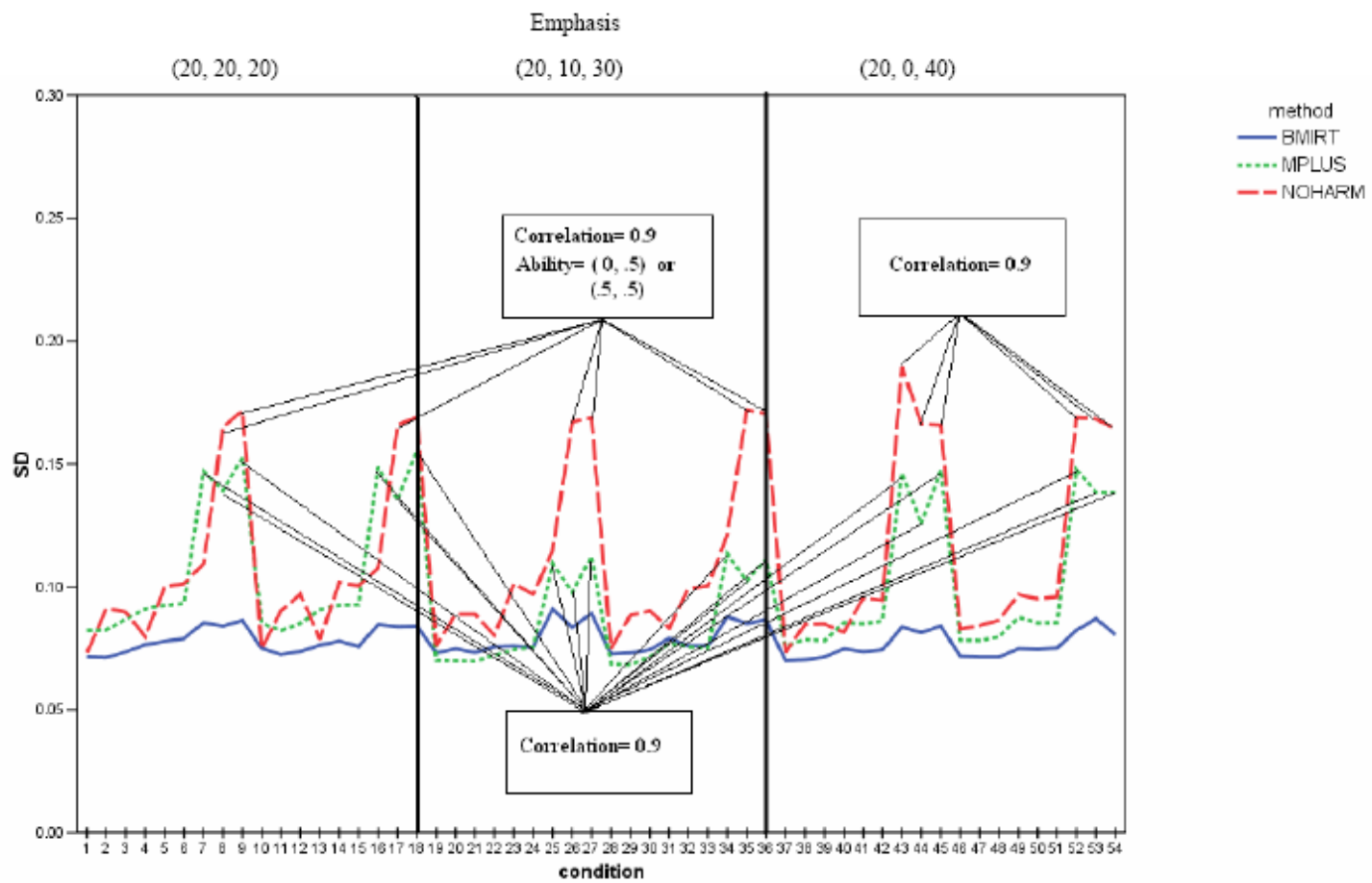


Figure 4-2  $SD$  of  $a_1$  from the three calibration methods  
under all 54 conditions

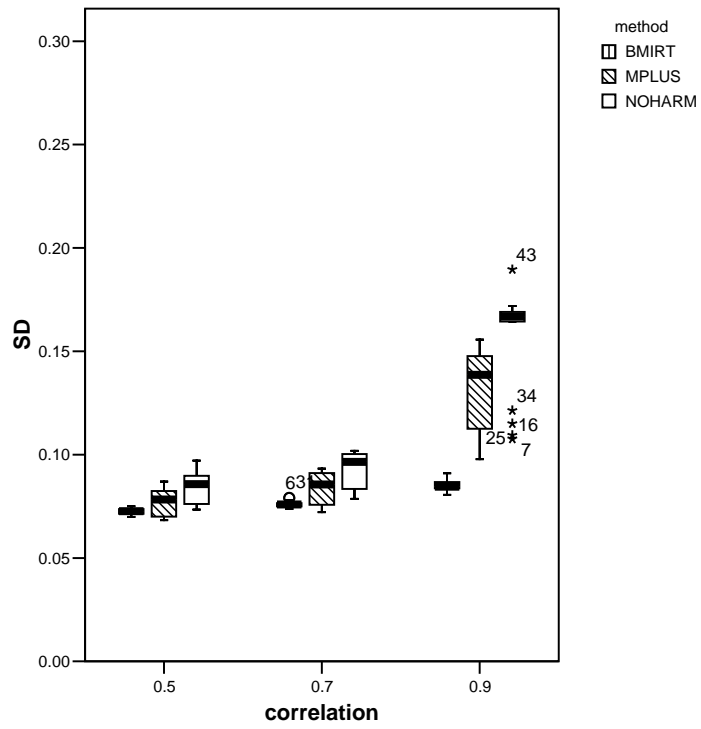
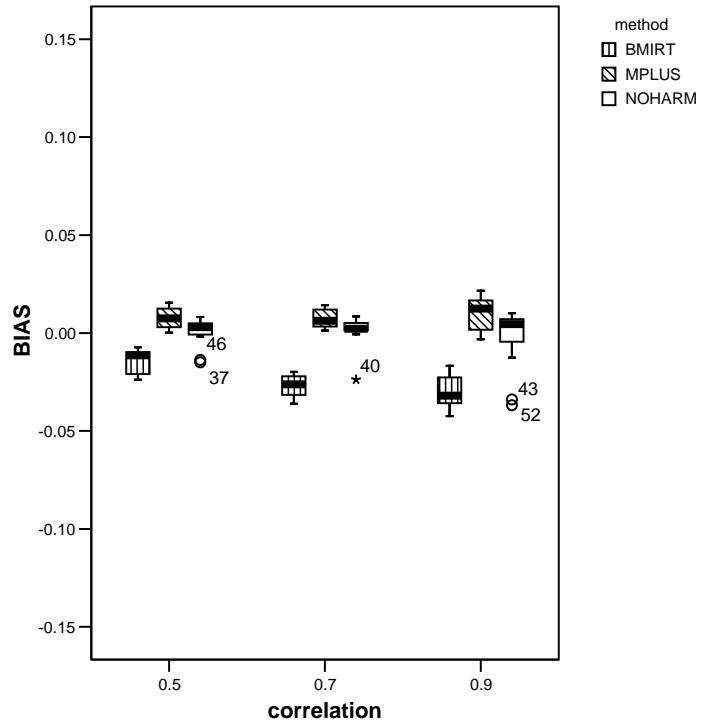


Figure 4-3 *BIAS* and *SD* of  $a_1$  under three structural orthogonality levels

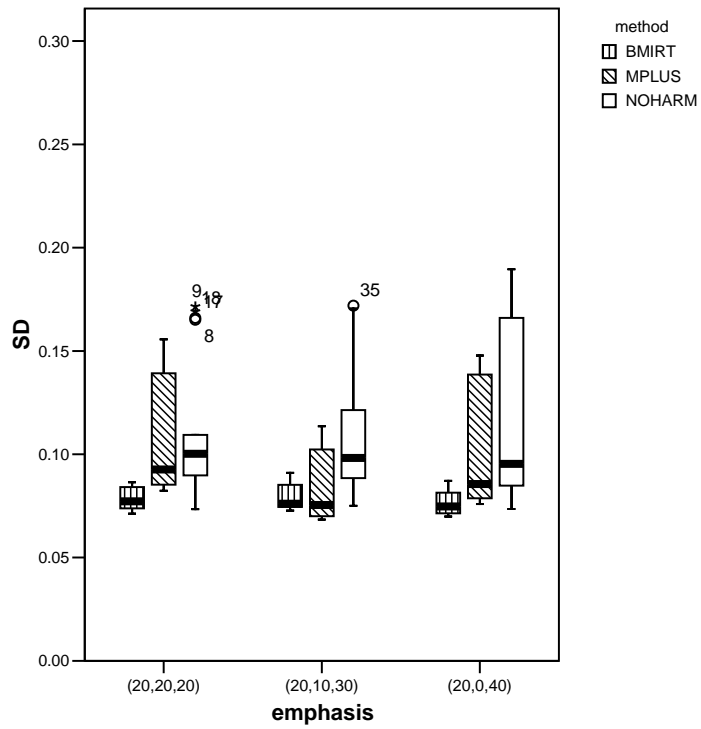
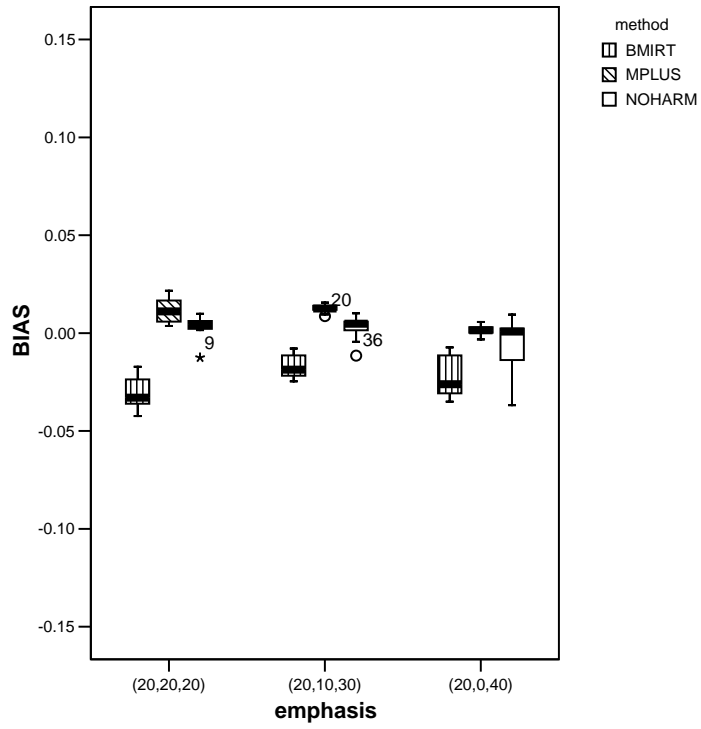


Figure 4-4 *BIAS* and *SD* of  $a_1$  under three structural equivalence levels

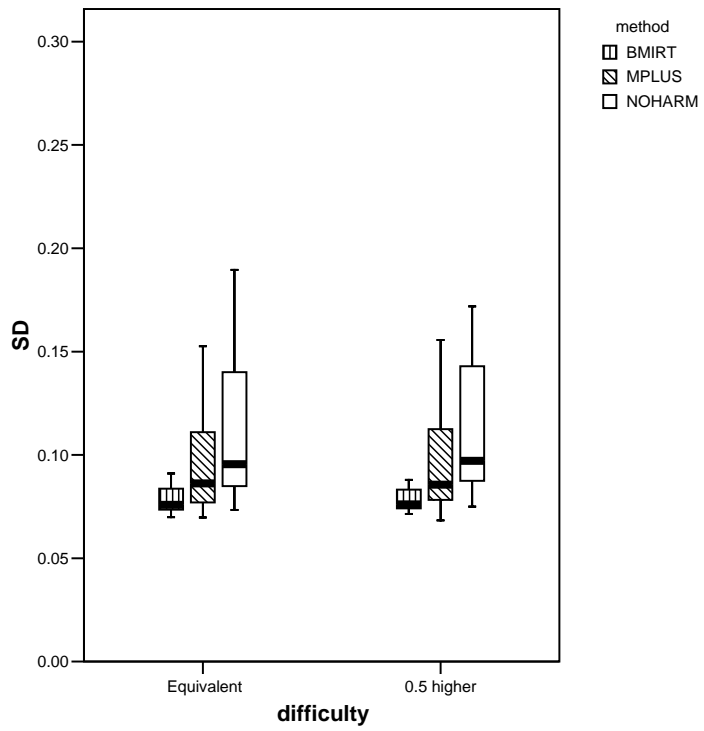
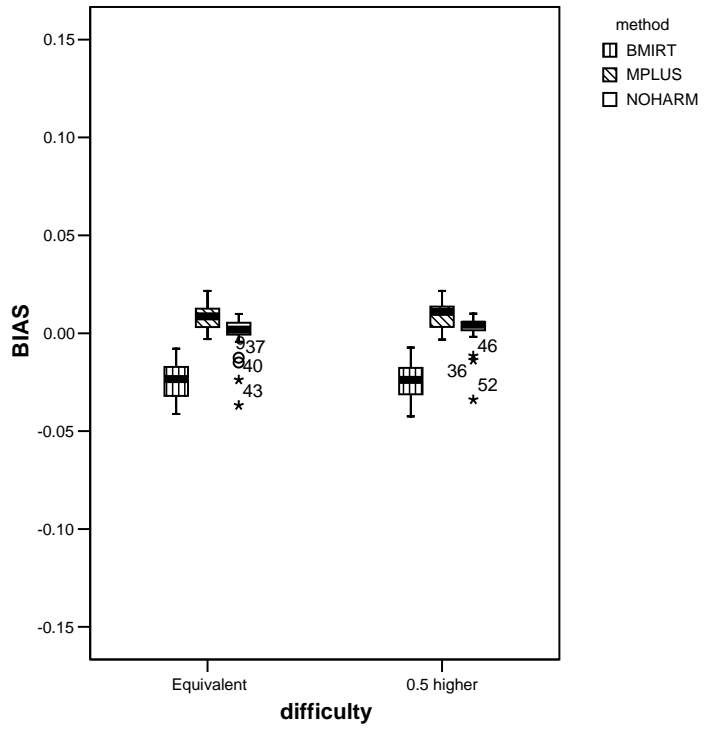


Figure 4-5 *BIAS* and *SD* of  $a_1$  under two item difficulty equivalence levels

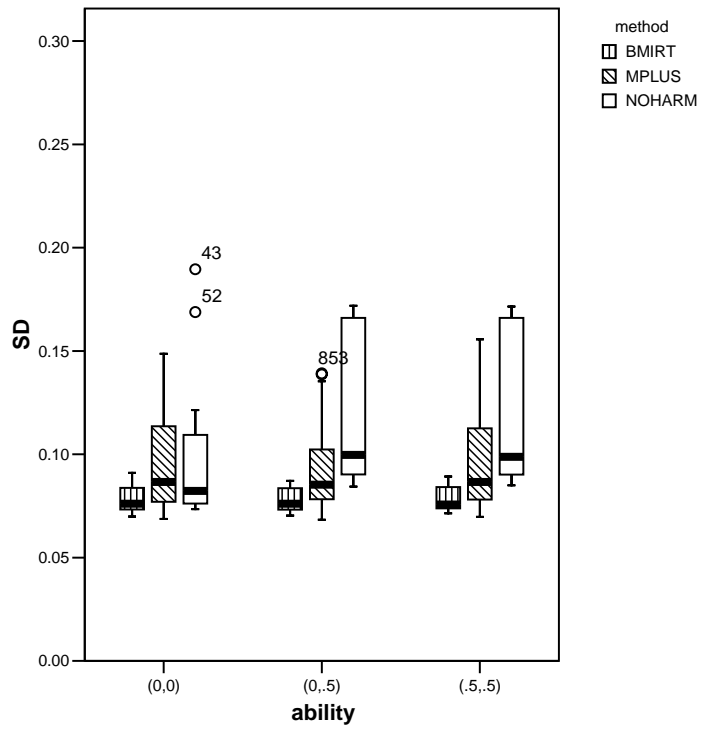
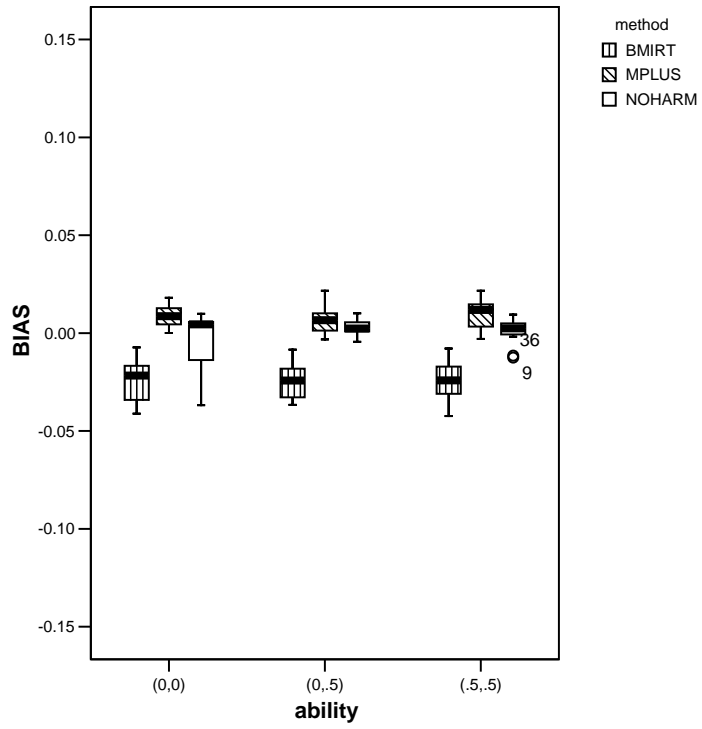


Figure 4-6 *BIAS* and *SD* of  $a_1$  under three examinee group equivalence levels

#### 4.1.2 The recovery of $a_2$

Figures 4-7 to 4-8 depict the *BIAS* and the *SD* of  $a_2$  from the three methods under all 54 conditions. Figures 4-9 to 4-12 depict the effect of the four manipulated factors on the *BIAS* and the *SD* of the estimate of  $a_2$  from the three methods. The notations in these figures are the same as those in Figures 4-1 to 4-6.

From Figures 4-7 to 4-12, it can be found that the three methods performed differently on the estimation of  $a_2$ . The following are the major findings.

- (1) In general, the *BIAS* of  $a_2$  was larger than that of  $a_1$  for all three methods under most conditions when Form 2 had more measurement emphasis on  $\theta_2$ . BMIRT tended to underestimate  $a_2$  under all conditions and the absolute magnitude of the *BIAS* was larger than that from the other two methods. Different from the estimate of  $a_1$ , the *BIAS* of  $a_2$  from NOHARM and Mplus were not comparable. The *BIAS* from BMIRT and NOHARM tended to fluctuate widely across conditions. In contrast, that from Mplus was more consistent, which indicates that the *BIAS* of  $a_2$  from Mplus was less affected by the manipulated factors than the other two methods.
- (2) The *SD* of  $a_2$  was comparable to that of  $a_1$  for all three methods under most conditions. The *SD* of  $a_2$  from NOHARM was generally larger than the other two methods and tended to fluctuate widely across conditions. The *SD* from Mplus also fluctuated across conditions but with smaller magnitude than that from NOHARM. Compared with the other two methods, the *SD* from BMIRT were more consistent across conditions.

- (3) When the correlation between  $\theta_1$  and  $\theta_2$  increased, the absolute magnitude of the *BIAS* from BMIRT increased. When the correlation was 0.9, the *BIAS* from NOHARM was much larger than when the correlation was 0.7 or 0.5. For all three methods, the *SD* of  $a_2$  increased as the correlation increased, especially when the correlation increased from 0.7 to 0.9.
- (4) When the emphasis on  $\theta_2$  in Form 2 increased, the absolute magnitude of the *BIAS* from BMIRT increased. When the “emphasis” was (20, 20, 20), the *BIAS* from Mplus was close to zero. However, when the “emphasis” was (20, 10, 30) or (20, 0, 40), Mplus tended to underestimate  $a_2$ . With respect to NOHARM, when the “emphasis” was (20, 0, 40), the *BIAS* was much larger than that when the “emphasis” was (20, 20, 20) or (20, 10, 30). Unlike the other two methods, NOHARM did not always underestimate  $a_2$ , sometimes it overestimated the parameter. An interesting finding was that when more emphasis was put on  $\theta_2$  in Form 2, the *SD* of  $a_2$  from Mplus decreased, which indicated that the estimate of the parameter became more stable.
- (5) When the two groups were not equivalent, NOHARM tended to underestimate  $a_2$  with larger *BIAS* and *SD* than when the two groups were equivalent.
- (6) There were some higher order interaction effects among the manipulated factors and the calibration methods on the *BIAS* of  $a_2$  from NOHARM. When the “emphasis” was (20, 20, 20) or (20, 10, 30) and the two groups were not equivalent, NOHARM tended to underestimate  $a_2$  when the



correlation was 0.9. When the “emphasis” was (20, 0, 40) and the two groups were not equivalent, NOHARM tended to underestimate  $a_2$  no matter what the correlation was.

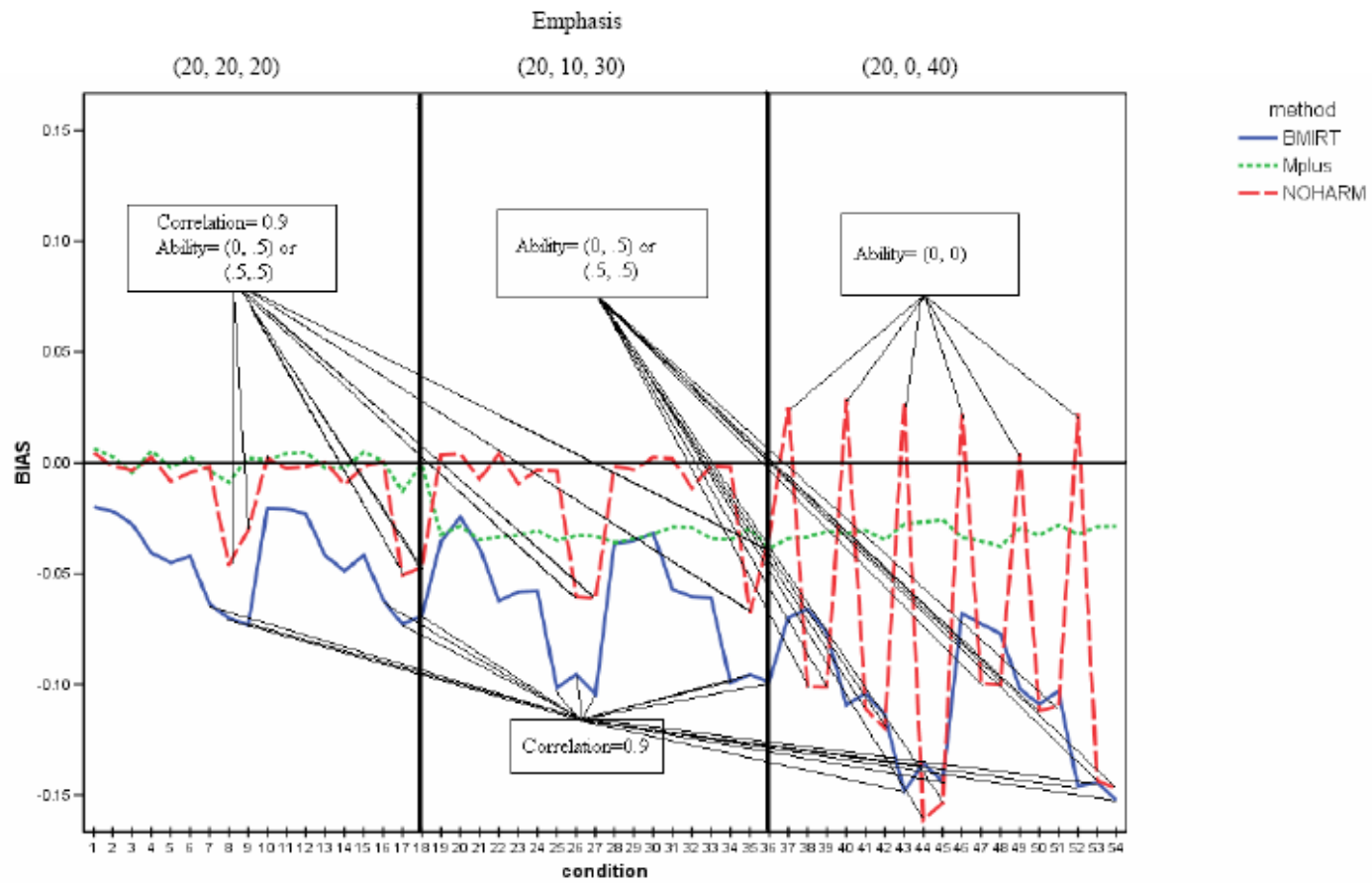


Figure 4-7 *BIAS* of  $a_2$  from the three calibration methods  
under all 54 conditions

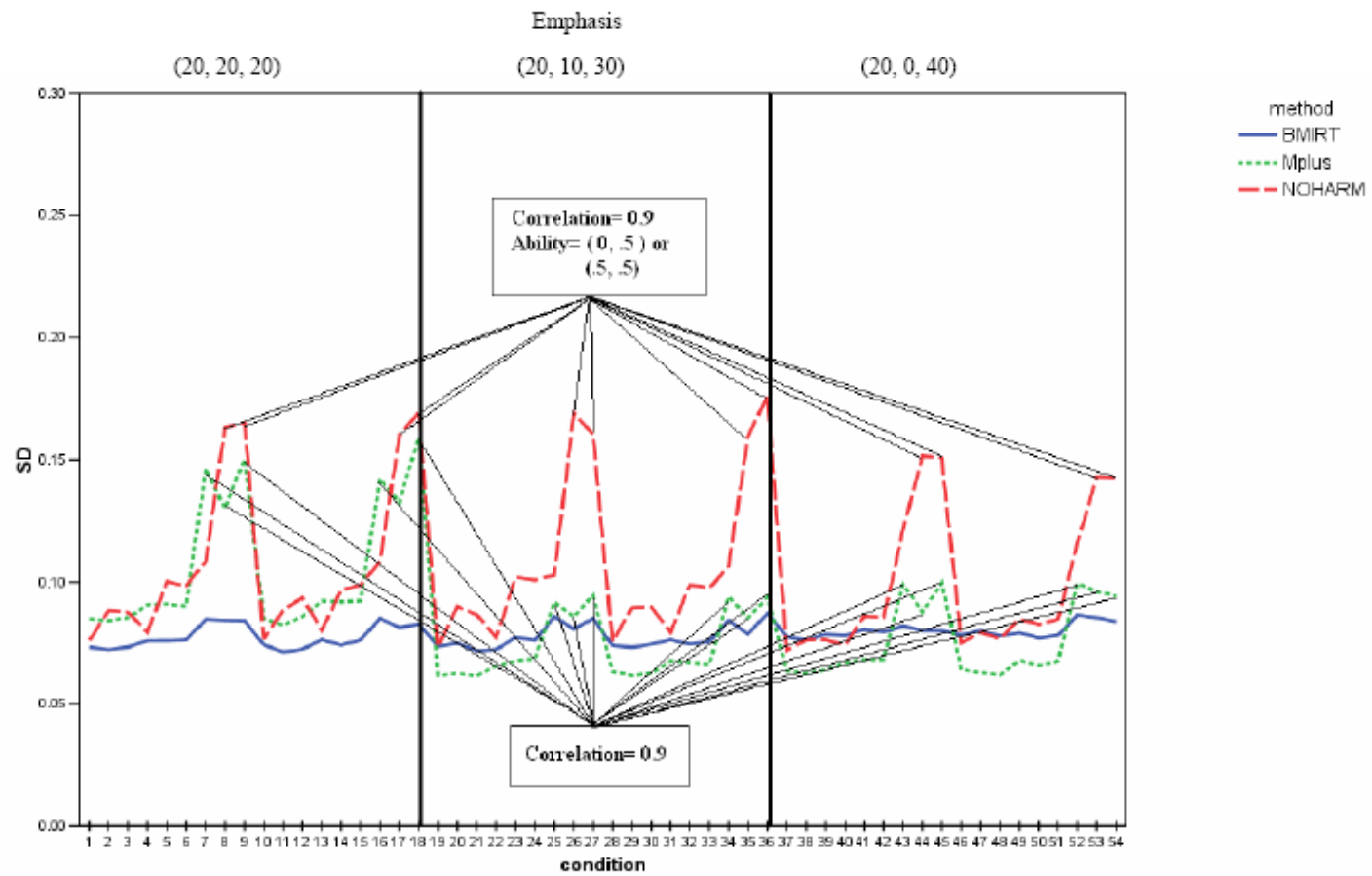


Figure 4-8  $SD$  of  $a_2$  from the three calibration methods  
under all 54 conditions

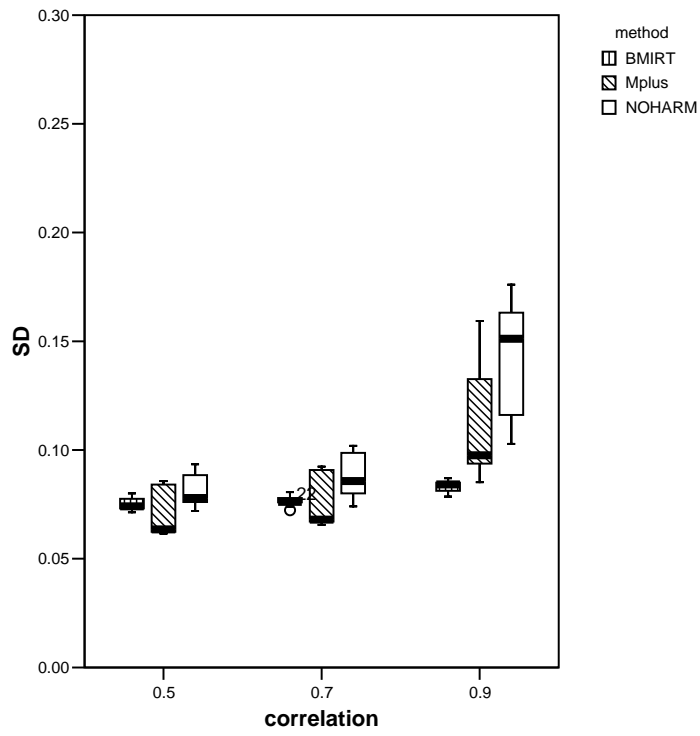
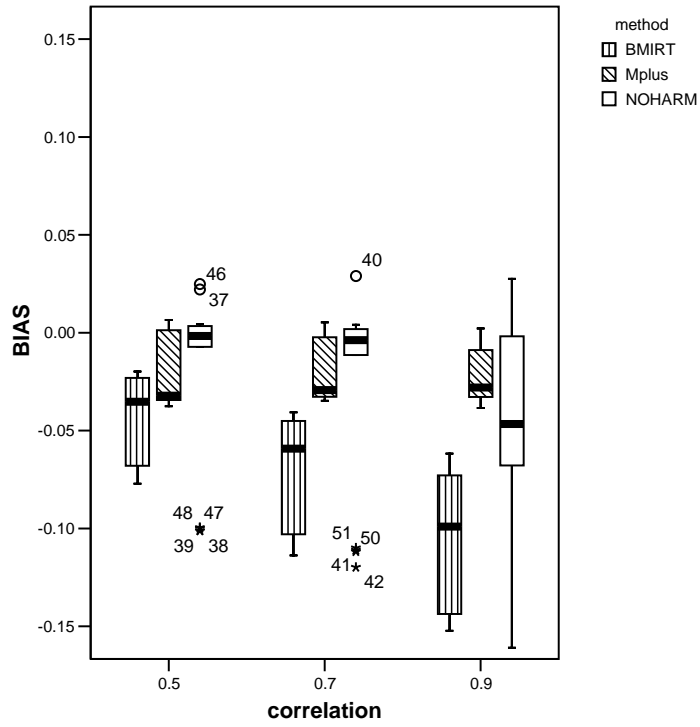


Figure 4-9 *BIAS* and *SD* of  $a_2$  under three structural orthogonality levels

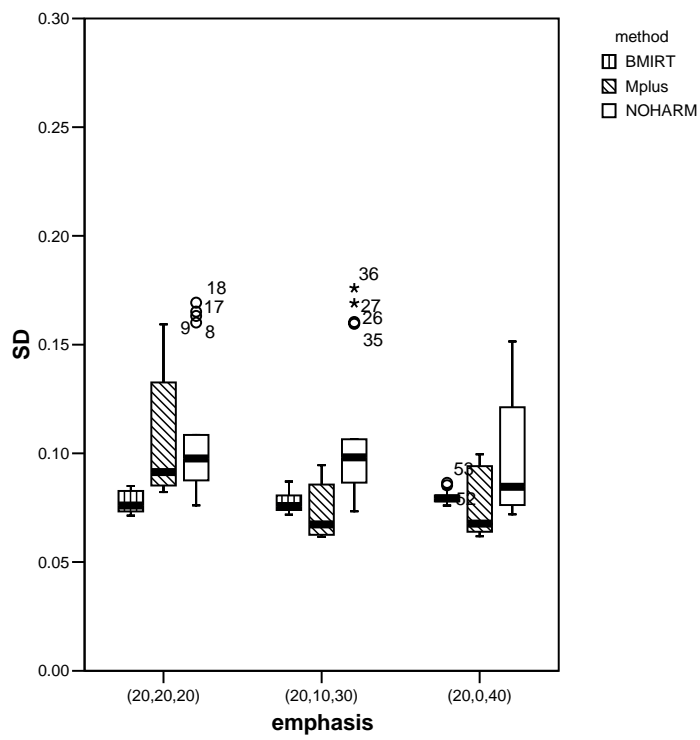
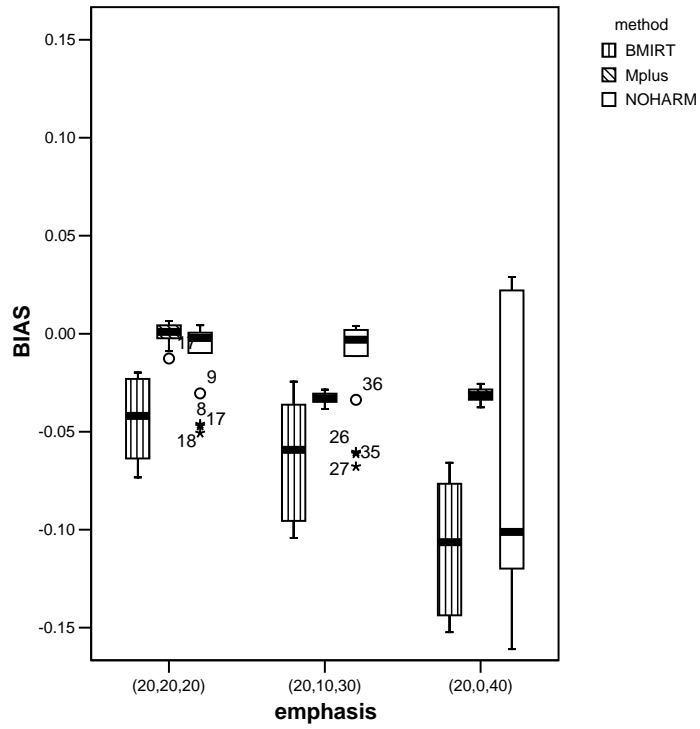


Figure 4-10 *BIAS* and *SD* of  $a_2$  under three structural equivalence levels

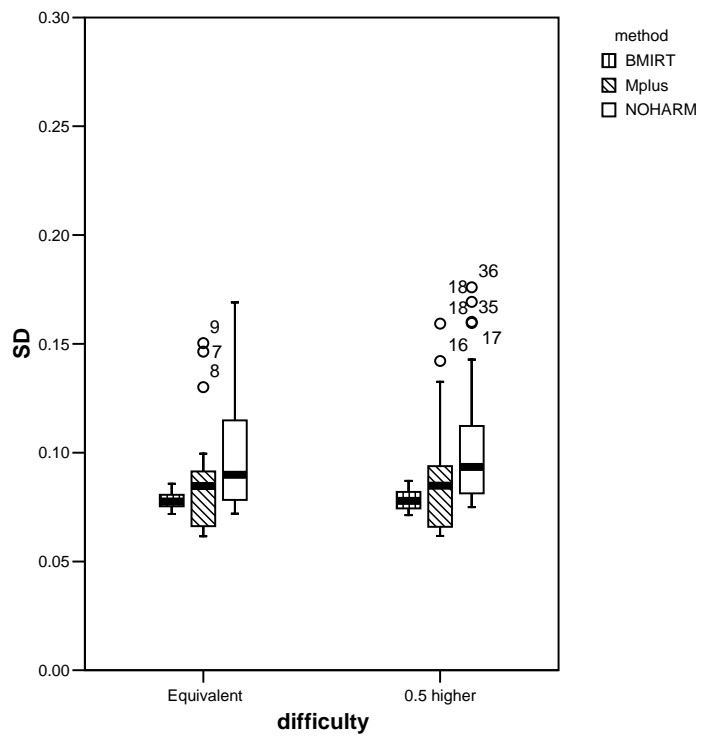
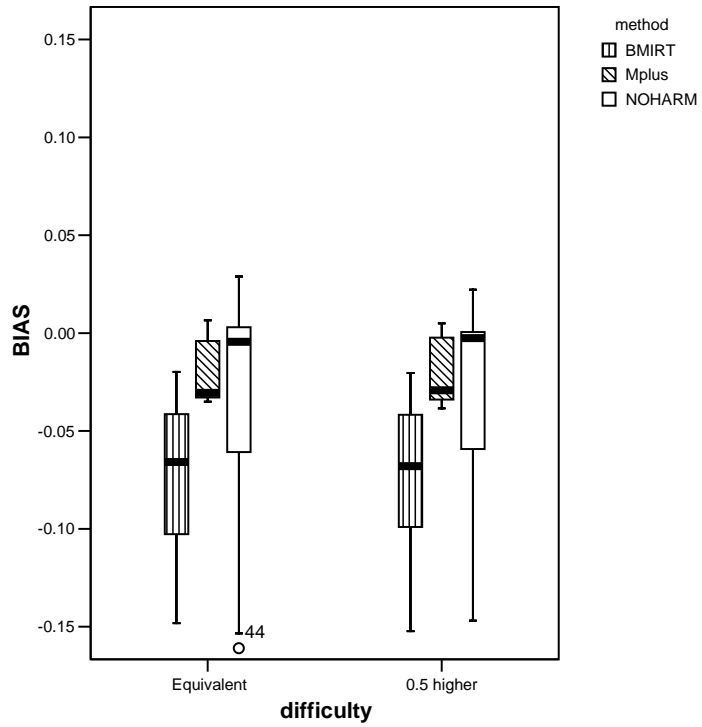


Figure 4-11 *BIAS* and *SD* of  $a_2$  under two item difficulty equivalence levels

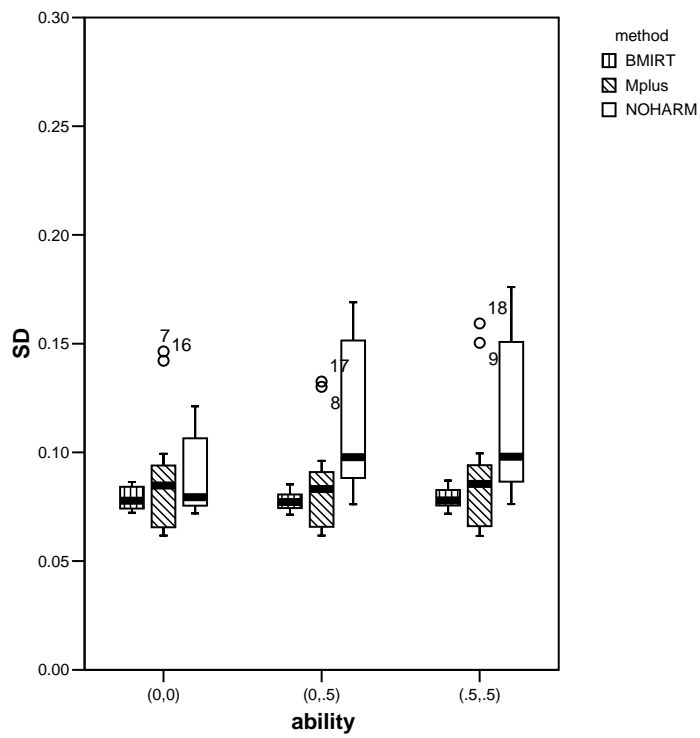
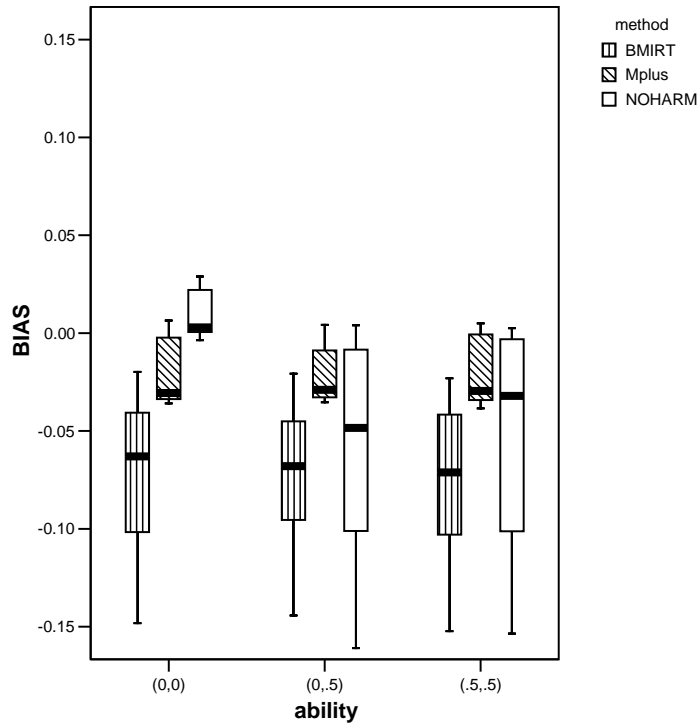


Figure 4-12 *BIAS* and *SD* of  $a_2$  under three examinee group equivalence levels

#### 4.1.3 The recovery of $d$

Figures 4-13 to 4-18 depict the *BIAS* and the *SD* of  $d$  from the three methods under all 54 conditions. The gray areas in Figure 4-13 represent all the conditions with the “difficulty” being “.5 higher”. Figures 4-9 to 4-12 depict the effect of the four manipulated factors on the *BIAS* and the *SD* of the estimates of  $a_2$  from the three methods.

From Figures 4-13 to 4-18, it can be found that the three methods performed differently on the estimate of  $d$ . The following are the major findings.

- (1) In general, the *BIAS* of  $d$  from Mplus and NOHARM were comparable and very close to zero under most conditions. BMIRT tended to overestimate  $d$  under most conditions, and the absolute magnitude of the *BIAS* was usually larger than that of the other two methods.
- (2) The three methods had comparable *SD* of  $d$ , except under some conditions the *SD* from Mplus became much larger than that from the other two methods.
- (3) When the correlation was 0.9, the *SD* of  $d$  from Mplus was much larger than when the correlation was 0.5 or 0.7.
- (4) When the emphasis on  $\theta_2$  in Form 2 increased, the *BIAS* of  $d$  from BMIRT increased.
- (5) When Form 2 was more difficult than Form 1, the *BIAS* of  $d$  from BMIRT was larger than when the two forms were equally difficult.
- (6) There was one higher order interaction effect. When the “ability” was (0, .5) and the “emphasis” was (20, 10, 30) or (20, 0, 40), NOHARM tended to overestimate  $d$  and the absolute magnitude of the *BIAS* was larger than that under the other conditions.



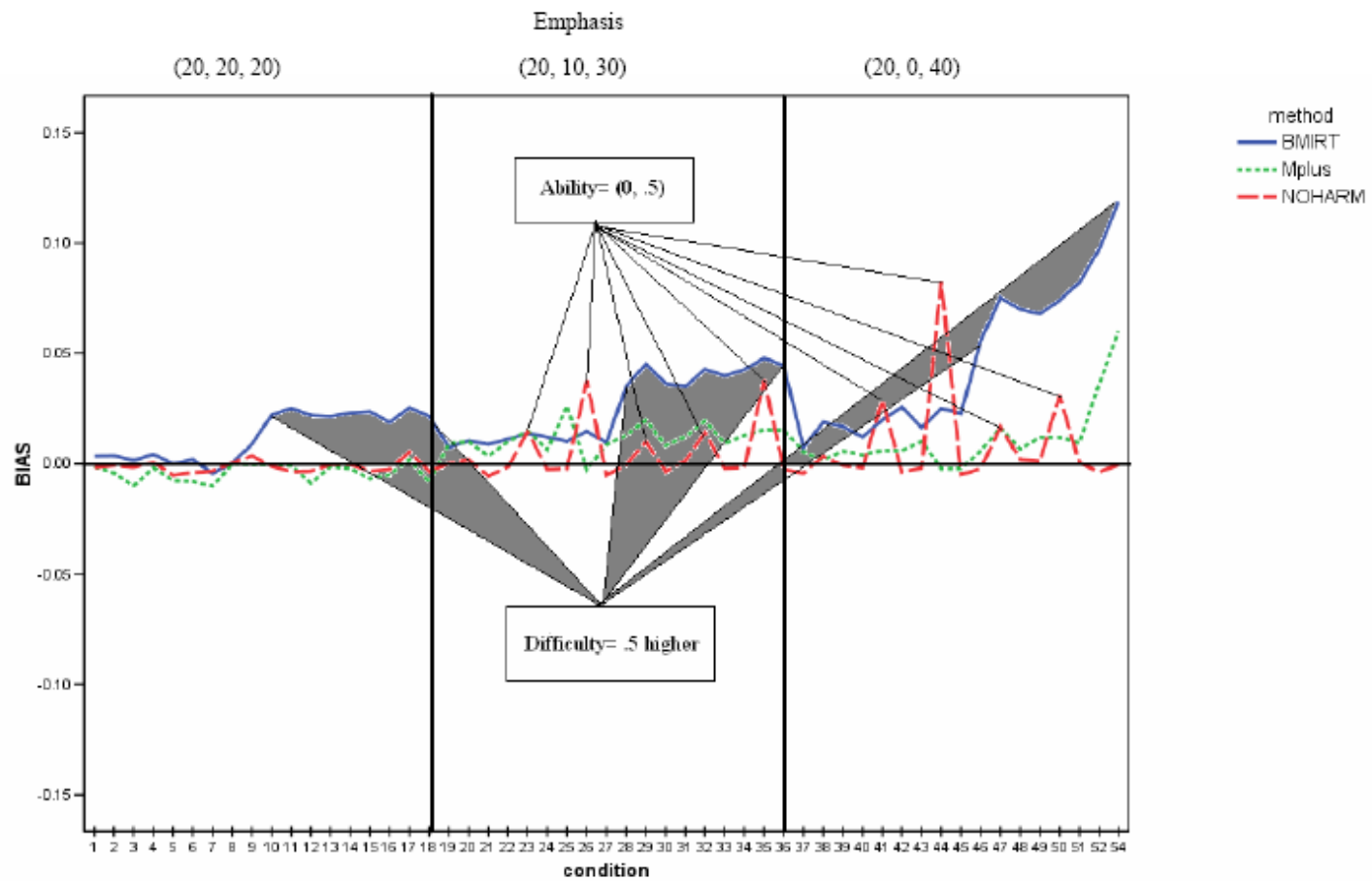


Figure 4-13 *BIAS* of *d* from the three calibration methods  
under all 54 conditions

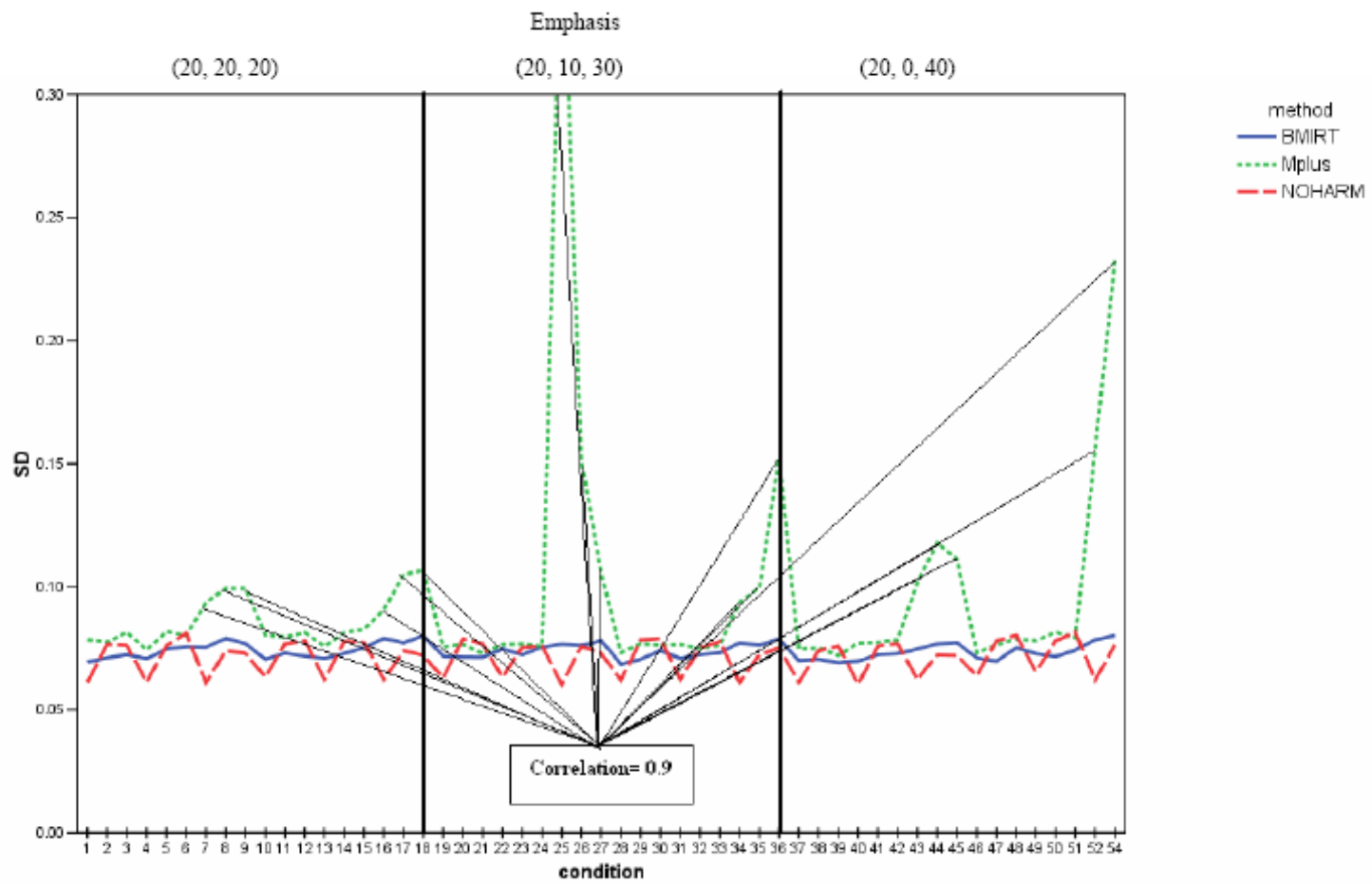


Figure 4-14 *SD* of *d* from the three calibration methods  
under all 54 conditions

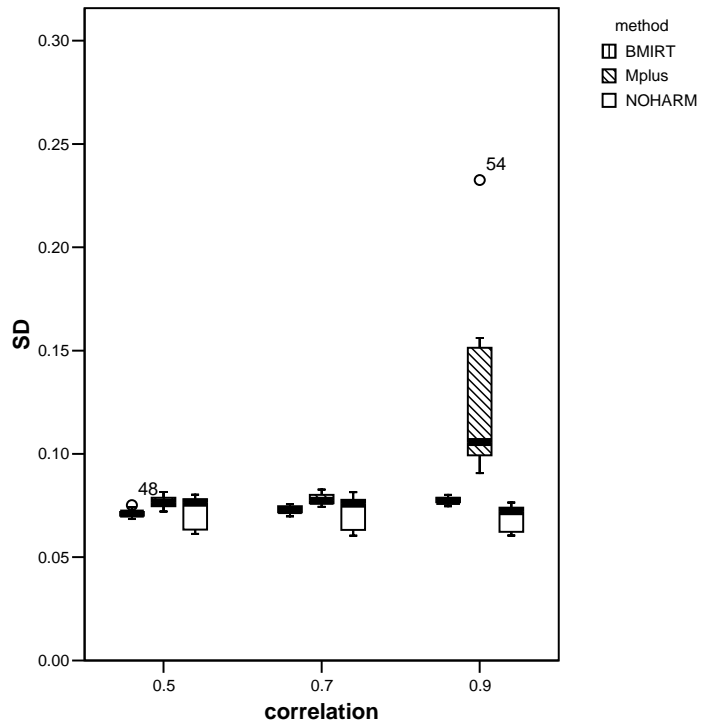
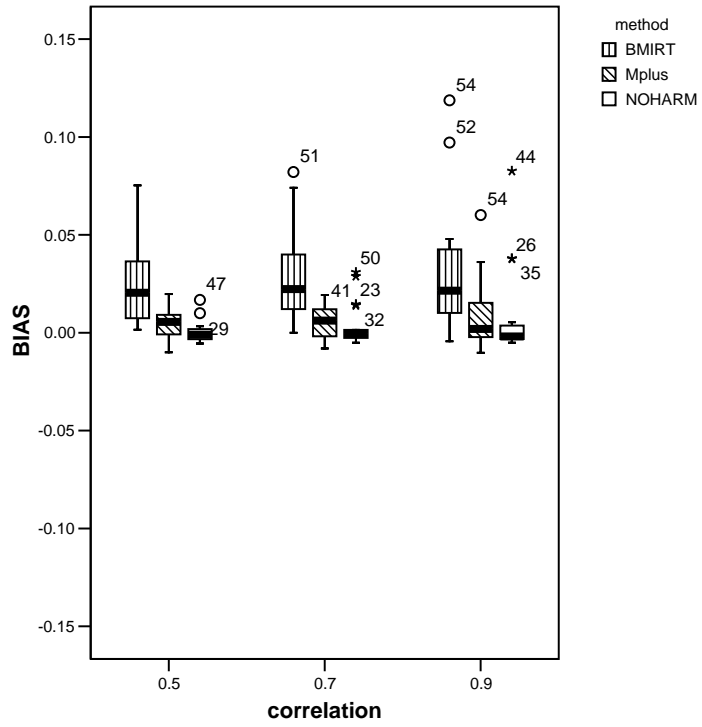


Figure 4-15 *BIAS* and *SD* of  $d$  under three structural orthogonality levels

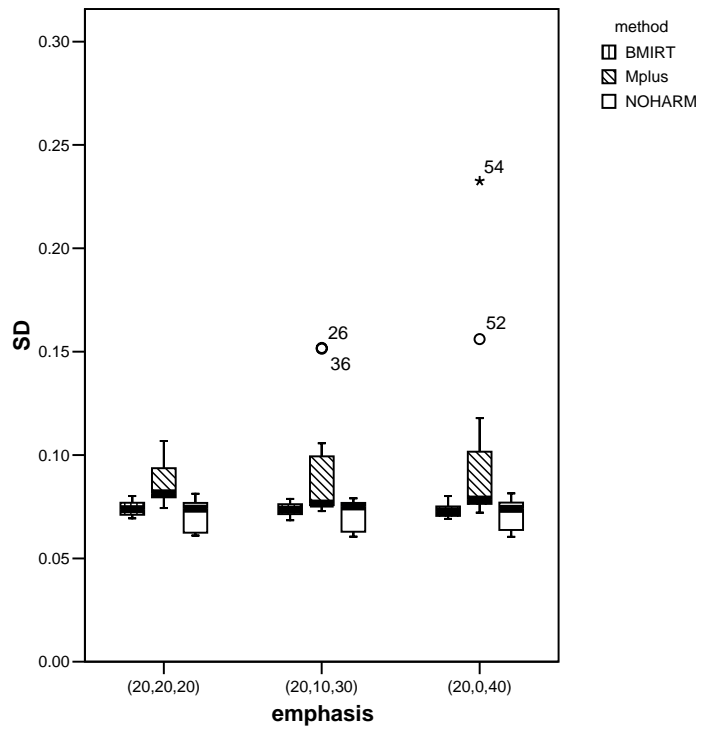
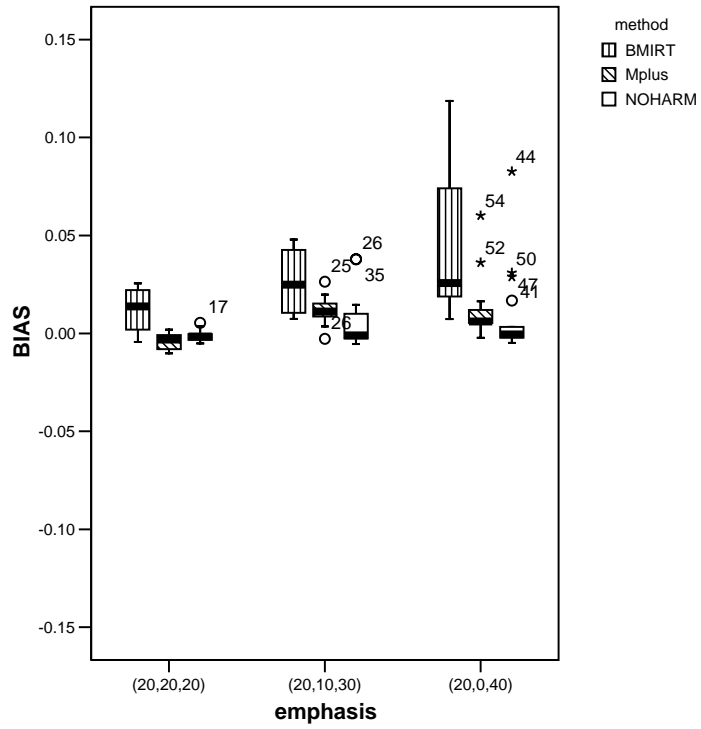


Figure 4-16 *BIAS* and *SD* of *d* under three structural equivalence levels

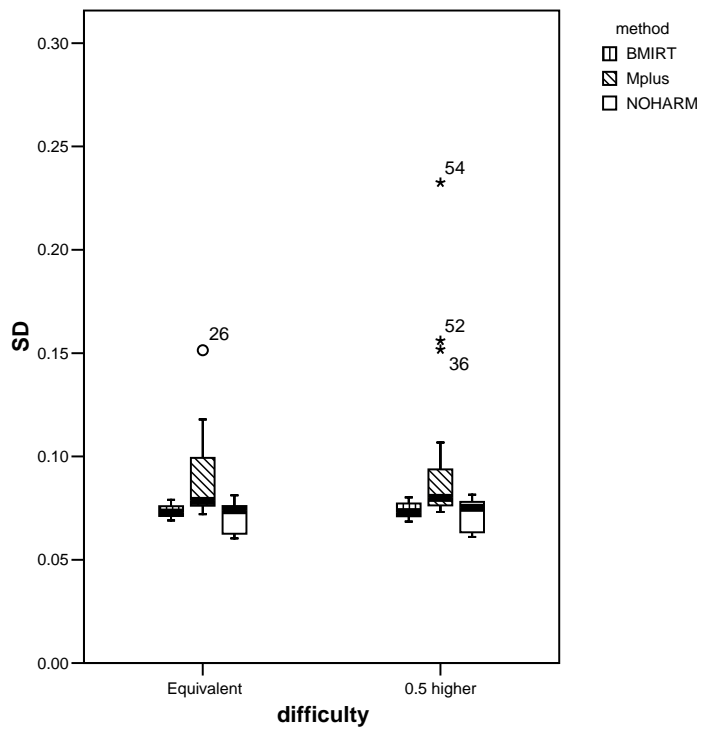
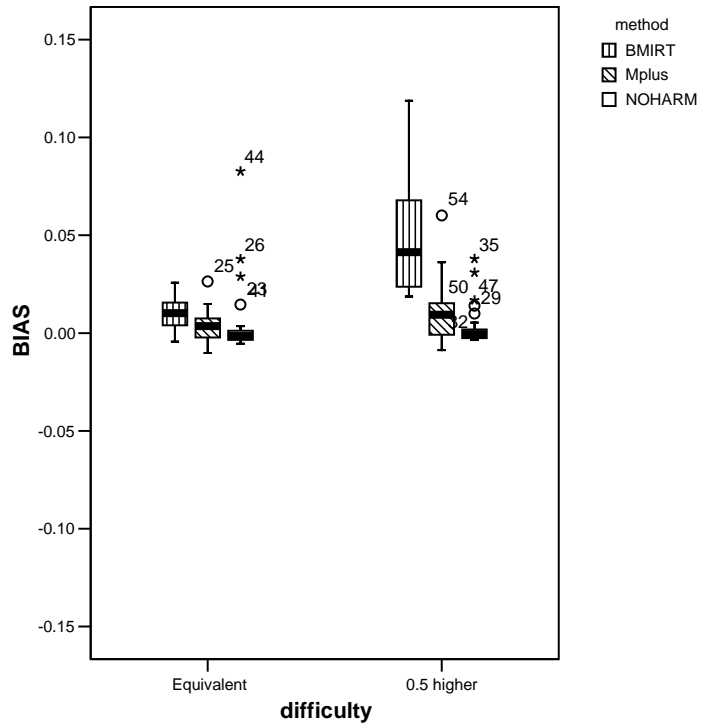


Figure 4-17 *BIAS* and *SD* of  $d$  under two item difficulty equivalence levels

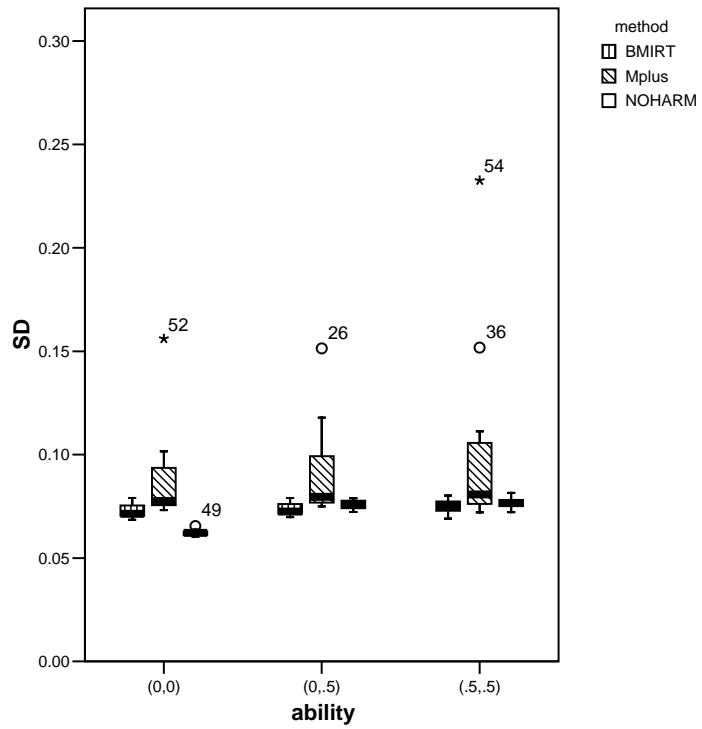
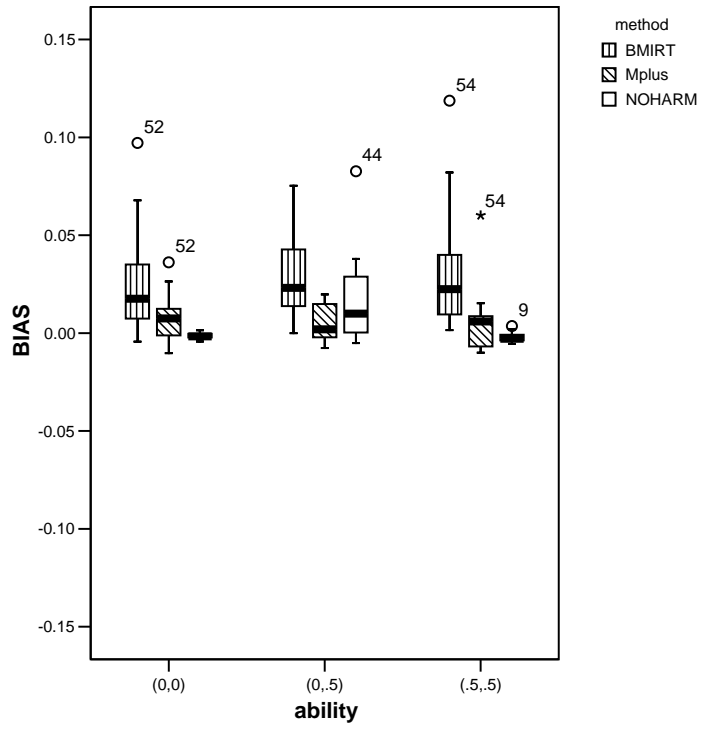


Figure 4-18 *BIAS* and *SD* of *d* under three examinee group equivalence levels

## 4.2 Estimate of true score of the examinees

The evaluation of the estimate of true scores was only available for two of the three multidimensional calibration methods (the concurrent MIRT method and the concurrent factor analysis method) and the unidimensional method (the concurrent UIRT method). Again, two criteria were used: *BIAS* measured the mean of the difference between the estimated true score and “true” true score of the examinees in each group; *SD* reflected the variability of the difference among the examinees in each group. More detailed information about the *BIAS* and the *SD* can be found in the tables in Appendix C.

### 4.2.1 The estimate of true scores in Group 1

Figures 4-19 and 4-20 depict the *BIAS* and the *SD* of the estimates of the true scores from the three methods in Group 1 under all 54 conditions. In the figures, the solid line represents the concurrent MIRT calibration method, which was carried out in the program BMIRT; the dotted line represents the concurrent factor analysis calibration method, which was carried out in the program Mplus; the dashed line represents the concurrent unidimensional IRT calibration method, which was carried out in the program BILOG. Again, the three methods are represented by the name of the programs that carried out the analysis. Each figure is split into three parts based on the three levels of the equivalence of the test structure between the two forms. In addition, the conditions with relative large *BIAS* are labeled with the corresponding value of the factor(s) that is(are) common to these conditions. Figures 4-21 to 4-24 depict the effect of the four manipulated factors on the *BIAS* and the *SD* of the estimate of true scores in Group 1 from the three methods.

From Figures 4-19 to 4-24, it can be found that the three methods performed

differently on the estimate of the true scores in Group 1. The following are the major findings.

- (1) In general, the *BIAS* of the estimate from all three methods was pretty small, with the magnitude less than 0.15 under most conditions, compared with the range of the true score, which was from 0 to 60.
- (2) The estimate of the true scores from BMIRT and BILOG were comparable and close to the true values under most conditions. The *BIAS* from Mplus tended to fluctuate across conditions and the absolute magnitude was usually larger than that from the other two methods.
- (3) The *SD* from BMIRT and BILOG was very comparable and consistent across conditions. The *SD* from Mplus was always larger than that from the other two methods and it tended to fluctuate across conditions.
- (4) When the “emphasis” was (20, 20, 20) or (20, 10, 30), Mplus tended to underestimate the true scores. When the “emphasis” was (20, 0, 40), however, it tended to overestimate the true scores. When the “emphasis” was (20, 0, 40), the *SD* from Mplus was much larger than when the “emphasis” was (20, 20, 20) or (20, 10, 30).
- (5) There were some higher order interaction effects among the factors and the calibration methods. When the “emphasis” was (20, 20, 20), the *BIAS* and the *SD* from Mplus was larger when the “ability” was (0, .5) or (.5, .5) than when the “ability” was (0, 0). When the “emphasis” was (20, 10, 30), the *BIAS* from Mplus increased with the increase of the correlation between  $\theta_1$  and  $\theta_2$ . An interesting finding was that when the “emphasis” was (20, 0, 40), the *SD* from Mplus decreased with the increase of the correlation between  $\theta_1$  and  $\theta_2$ .



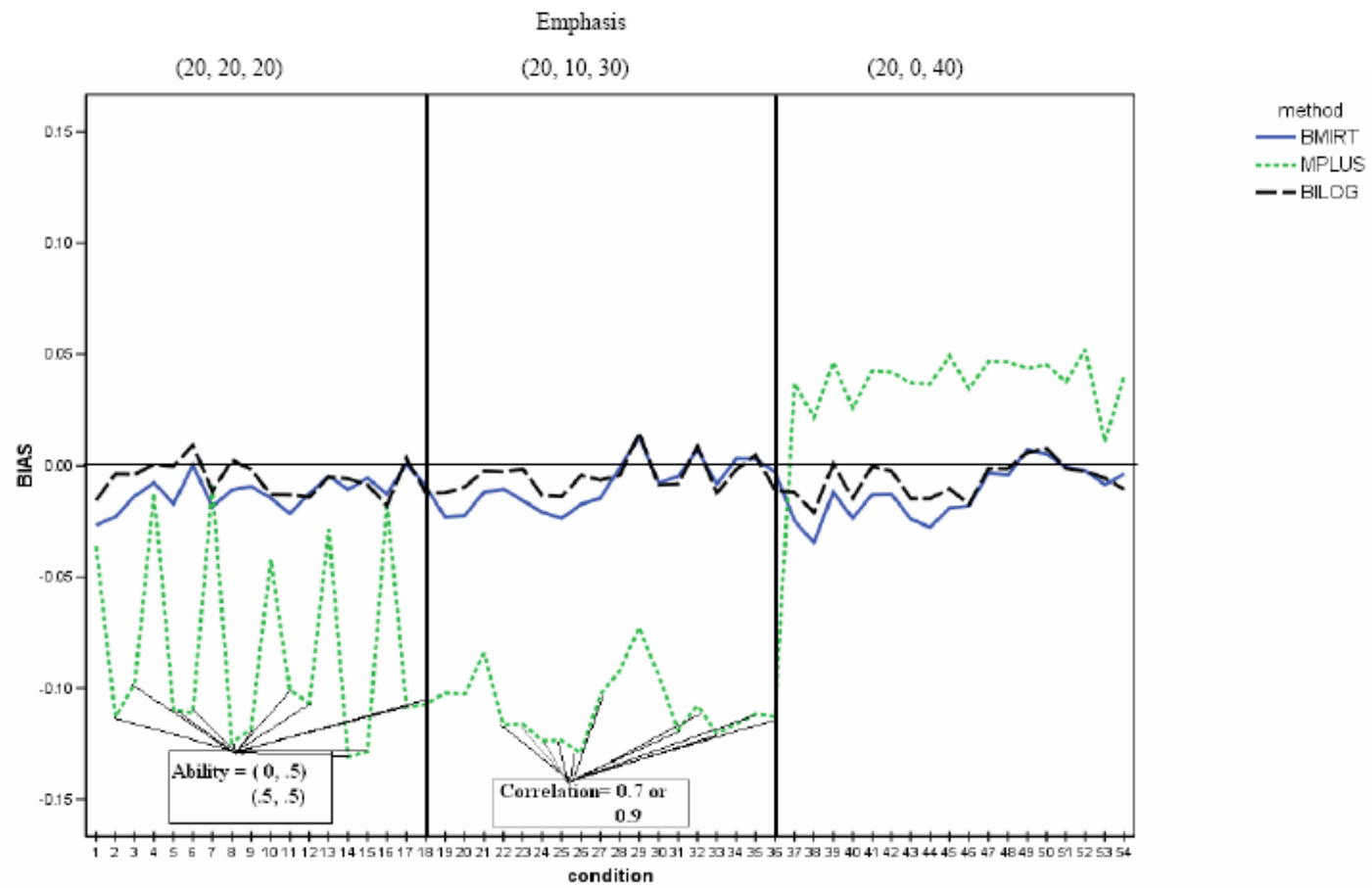


Figure 4-19 BIAS of the true score estimation in Group 1

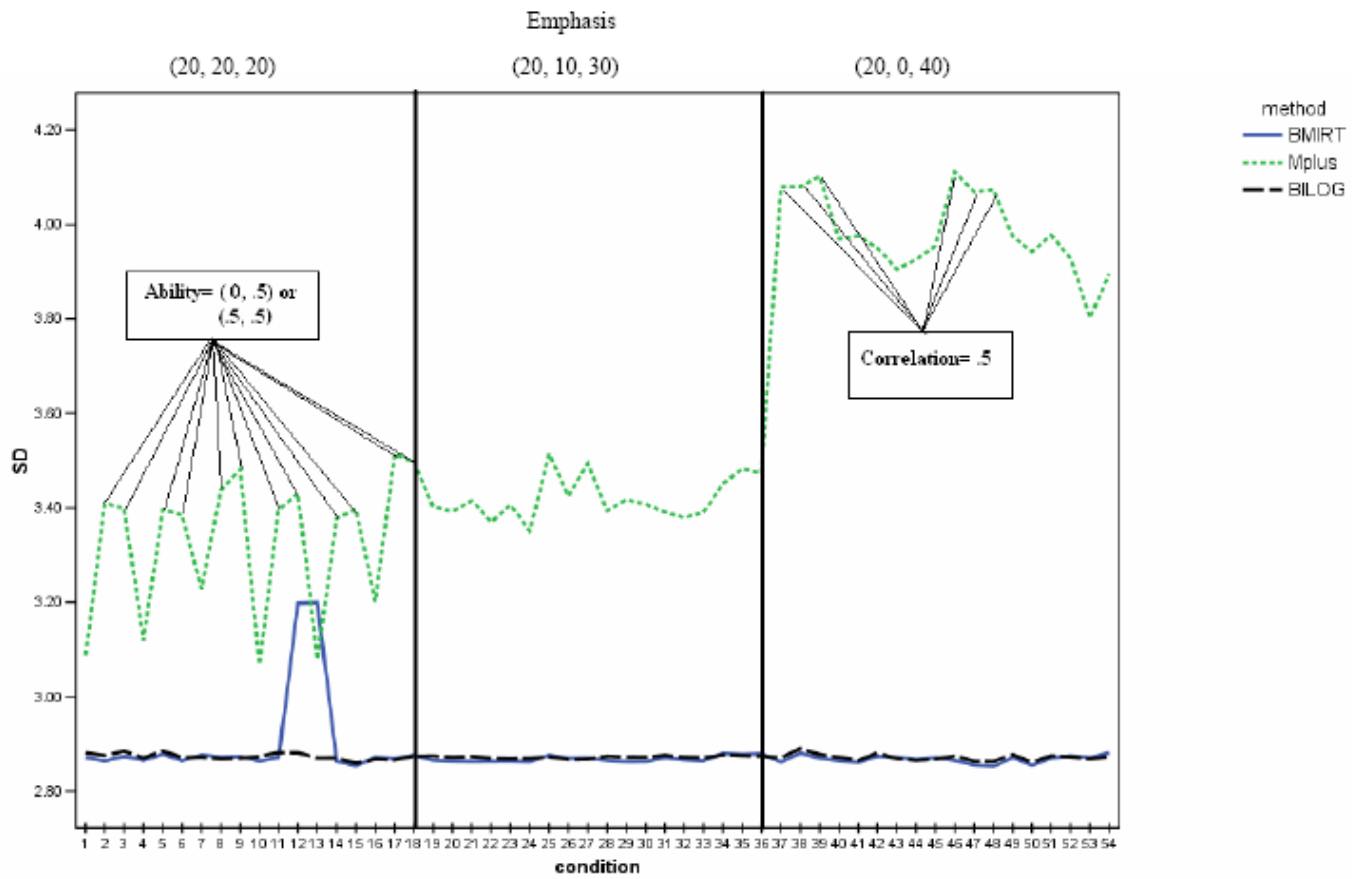


Figure 4-20 SD of the true score estimation in Group 1

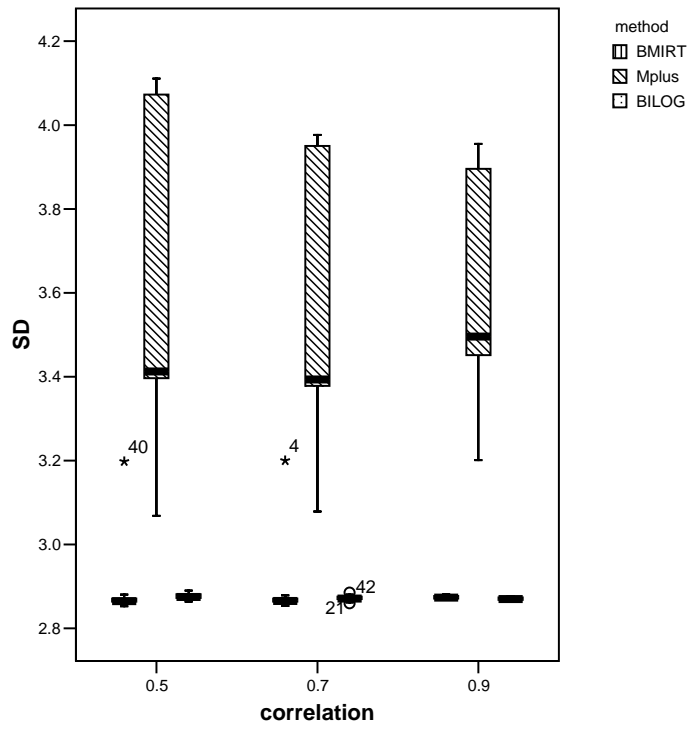
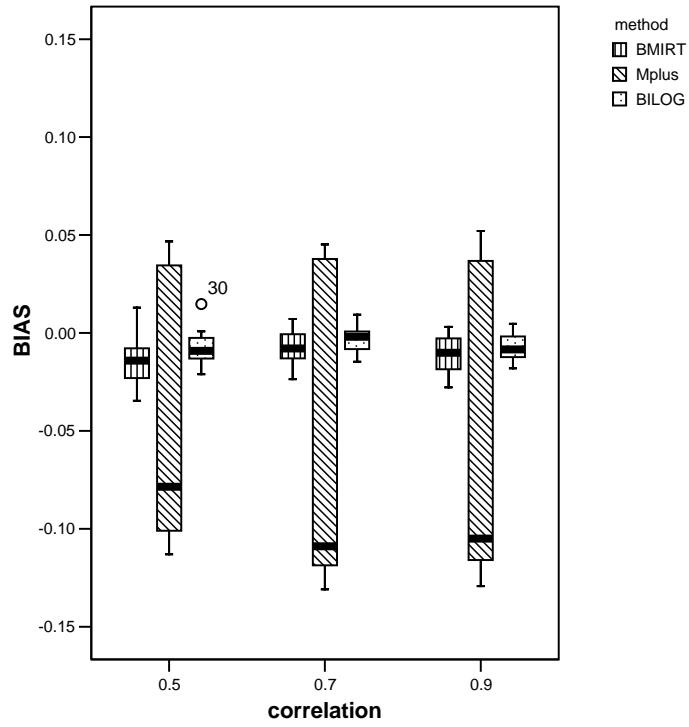


Figure 4-21 *BIAS* and *SD* of estimation of true score in Group 1 under three structural orthogonality levels

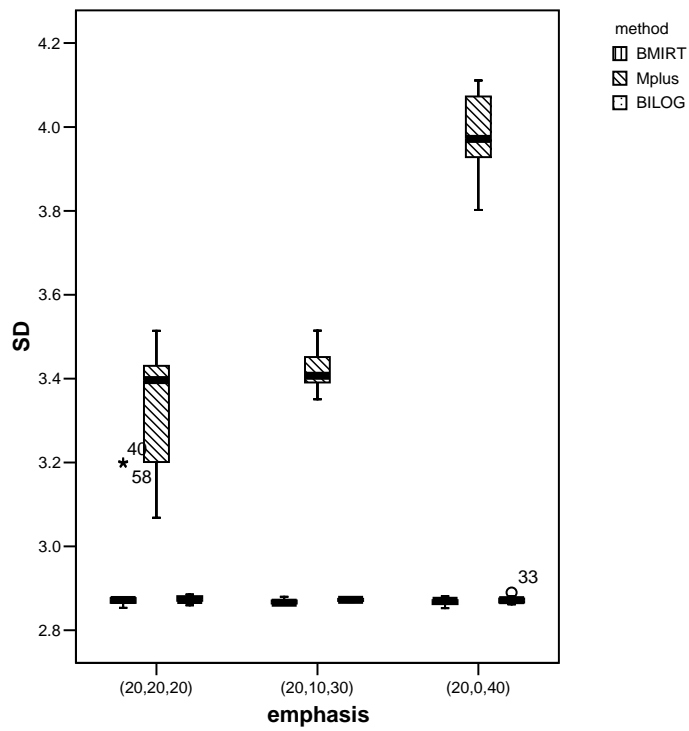
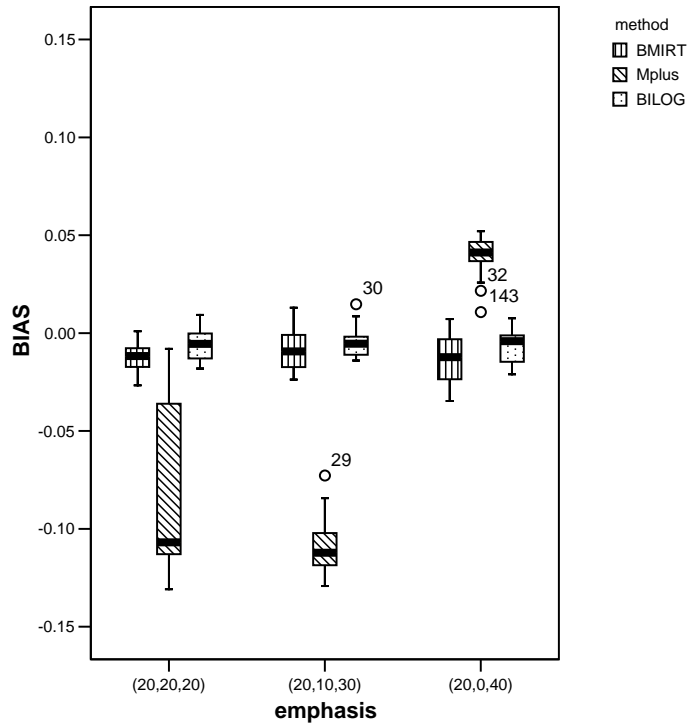


Figure 4-22 *BIAS* and *SD* of estimation of true score in Group 1 under three structural equivalence levels

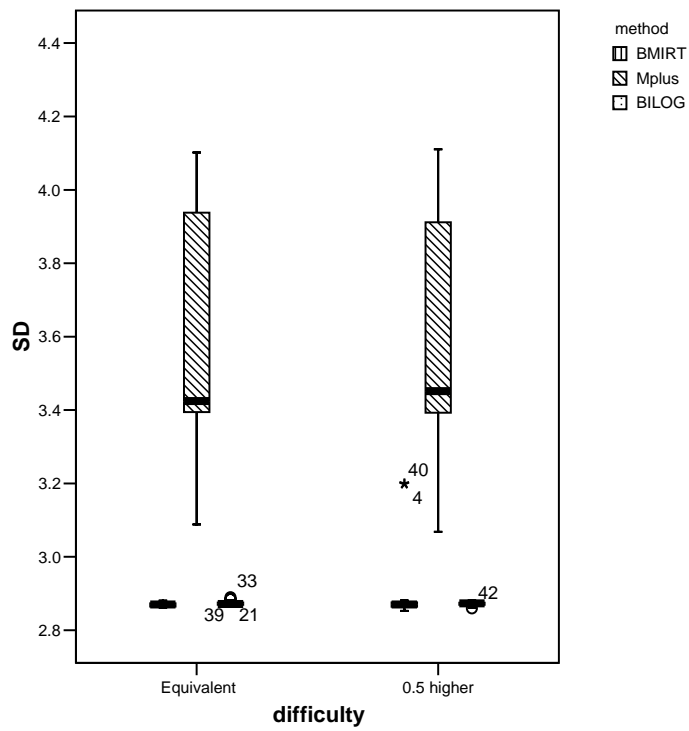
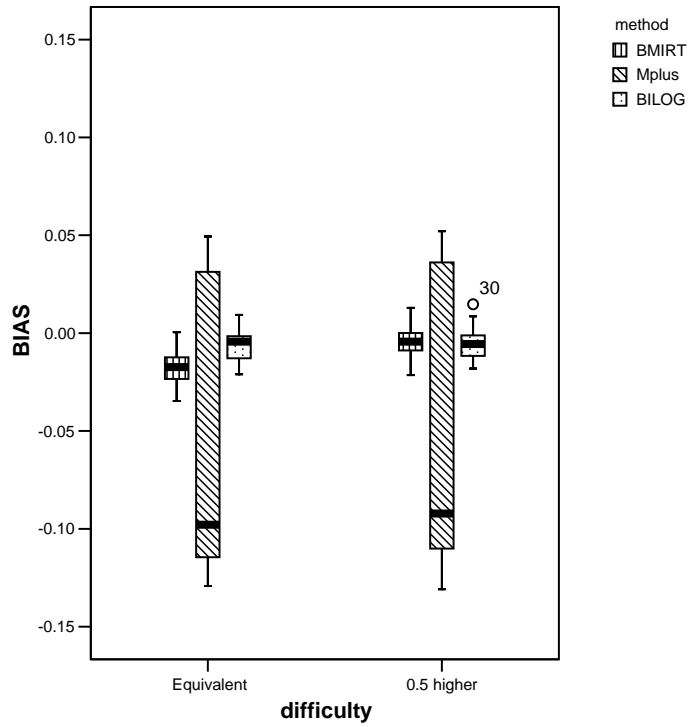


Figure 4-23 *BIAS* and *SD* of estimation of true score in Group 1 under two item difficulty equivalence levels

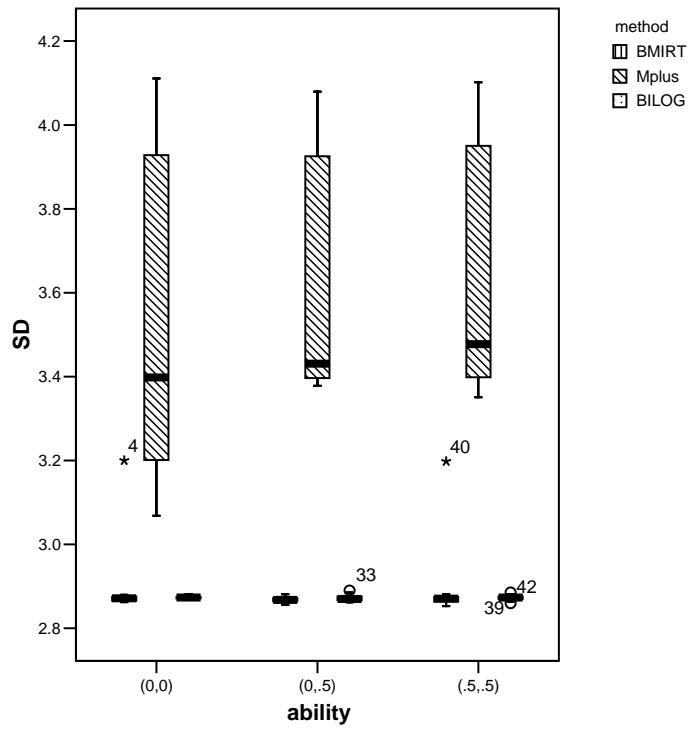
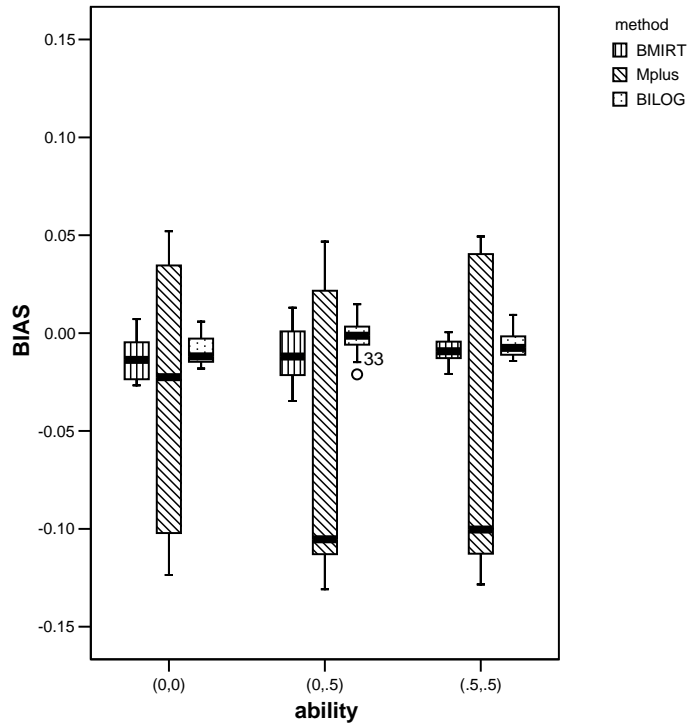


Figure 4-24 *BIAS* and *SD* of estimation of true score in Group 1 under three examinee group equivalence levels

#### 4.2.2 The true score estimation in Group 2

Figures 4-25 and 4-26 depict the *BIAS* and the *SD* of the estimates of the true scores from the three methods in Group 2 under all 54 conditions. Figures 4-27 to 4-30 depict the effect of the four manipulated factors on the *BIAS* and the *SD* of the estimate from the three methods.

From Figures 4-25 to 4-30, it can be found that the three methods performed differently on the estimate of the true scores in Group 2. The following are the major findings.

- (1) As in Group 1, in general, the *BIAS* of the estimates from all three methods was pretty small, with the magnitude less than 0.15 under most conditions, compared with the range of the true score, which was from 0 to 60.
- (2) The *BIAS* from BILOG was usually smaller than that from the other two methods. The *BIAS* from BMIRT and BILOG tended to fluctuate across conditions in the same pattern and the magnitude was usually much larger than that in Group 1. Mplus tended to overestimate the true score in Group 2 under most conditions. As in Group 1, the *BIAS* from Mplus in Group 2 was different from the other two methods. However, it was not always larger than that from the other two methods; sometimes it was smaller.
- (3) The *SD* from BMIRT and BILOG was very comparable and consistent across conditions. The absolute magnitude of *SD* from the two methods was much smaller than that from Mplus under all conditions. The *SD* from Mplus in Group 2 was usually smaller than that in Group 1 and it tended to fluctuate across conditions.
- (4) When the correlation between  $\theta_1$  and  $\theta_2$  was 0.9, Mplus tended to

overestimate the true scores to a slightly larger extent than that when the correlation was 0.5 or 0.7, and the *SD* of the estimate was also slightly larger.

- (5) When the “emphasis” was (20, 0, 40), the *BIAS* from BMIRT and BILOG was slightly smaller than when the “emphasis” was (20, 20, 20) or (20, 10, 30). In contrast, the *BIAS* from Mplus was slightly larger when the “emphasis” was (20, 0, 40) than when the “emphasis” was (20, 20, 20) or (20, 10, 30).
- (6) When the two forms had equivalent item difficulty level, all three methods tended to overestimate the true scores. But when Form 2 was more difficult than Form 1, BMIRT and BILOG, on average, no longer overestimated the true scores. Such change was not found in the estimate from Mplus.
- (7) When the two groups were equivalent, BMIRT and BILOG tended to underestimate the true scores under most conditions; but when Group 2 had higher ability than Group 1, the two methods tended to overestimate the true scores in Group 2. In addition, the absolute magnitude of the *BIAS* was larger when the “ability” was (.5, .5) than when the “ability” was (0, .5).
- (8) There were some higher order interaction effects among the factors and the calibration methods. When the “emphasis” was (20, 20, 20), the *SD* from Mplus was larger when the “ability” was (0, .5) than when the “ability” was (0, 0) or (.5, .5). An interesting finding was that when the “ability” was (0, 0), the *SD* from Mplus was larger than when the “ability” was (0, .5) or (.5, .5).



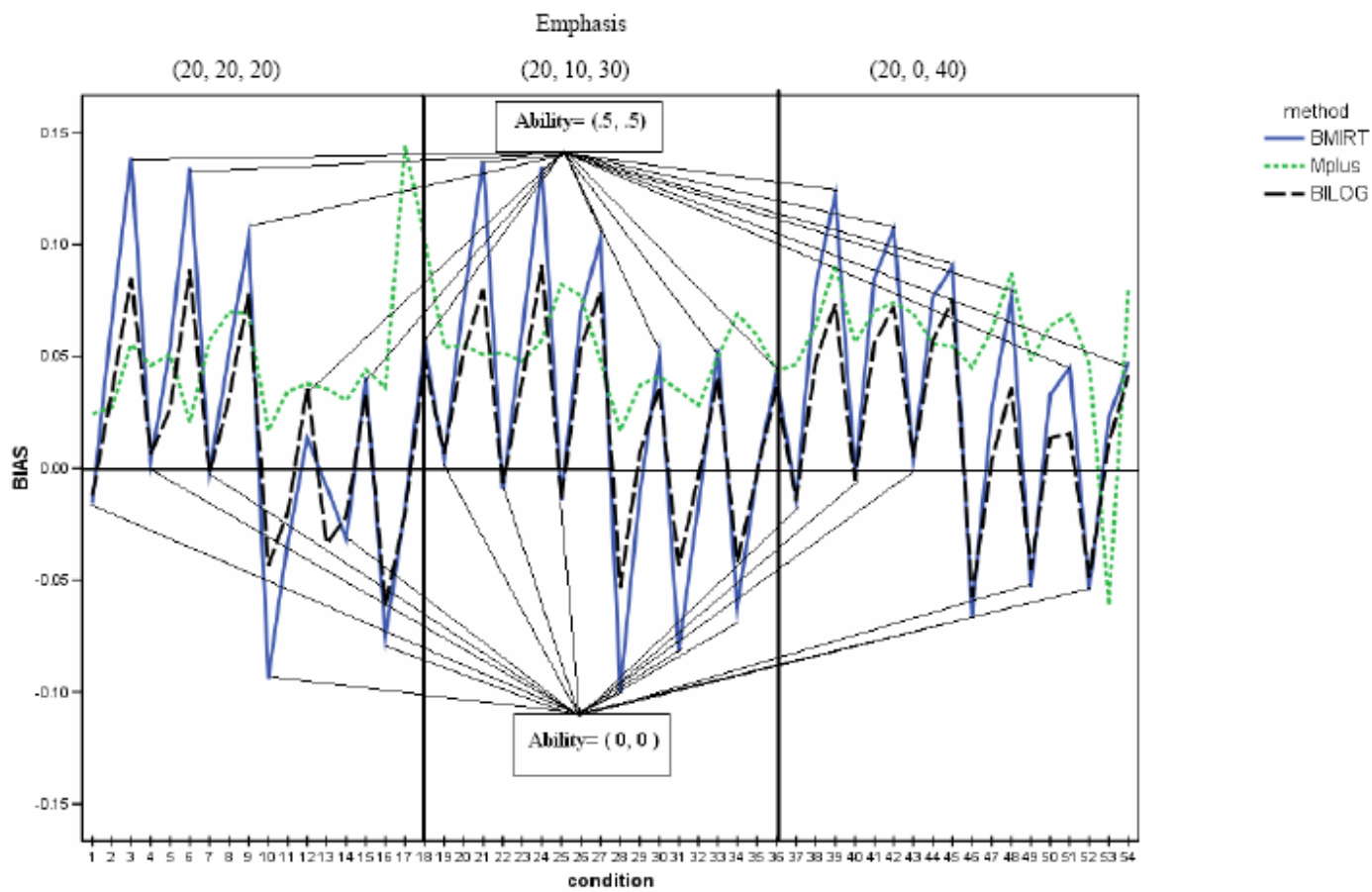


Figure 4-25 *BIAS* of the true score estimation in Group 2

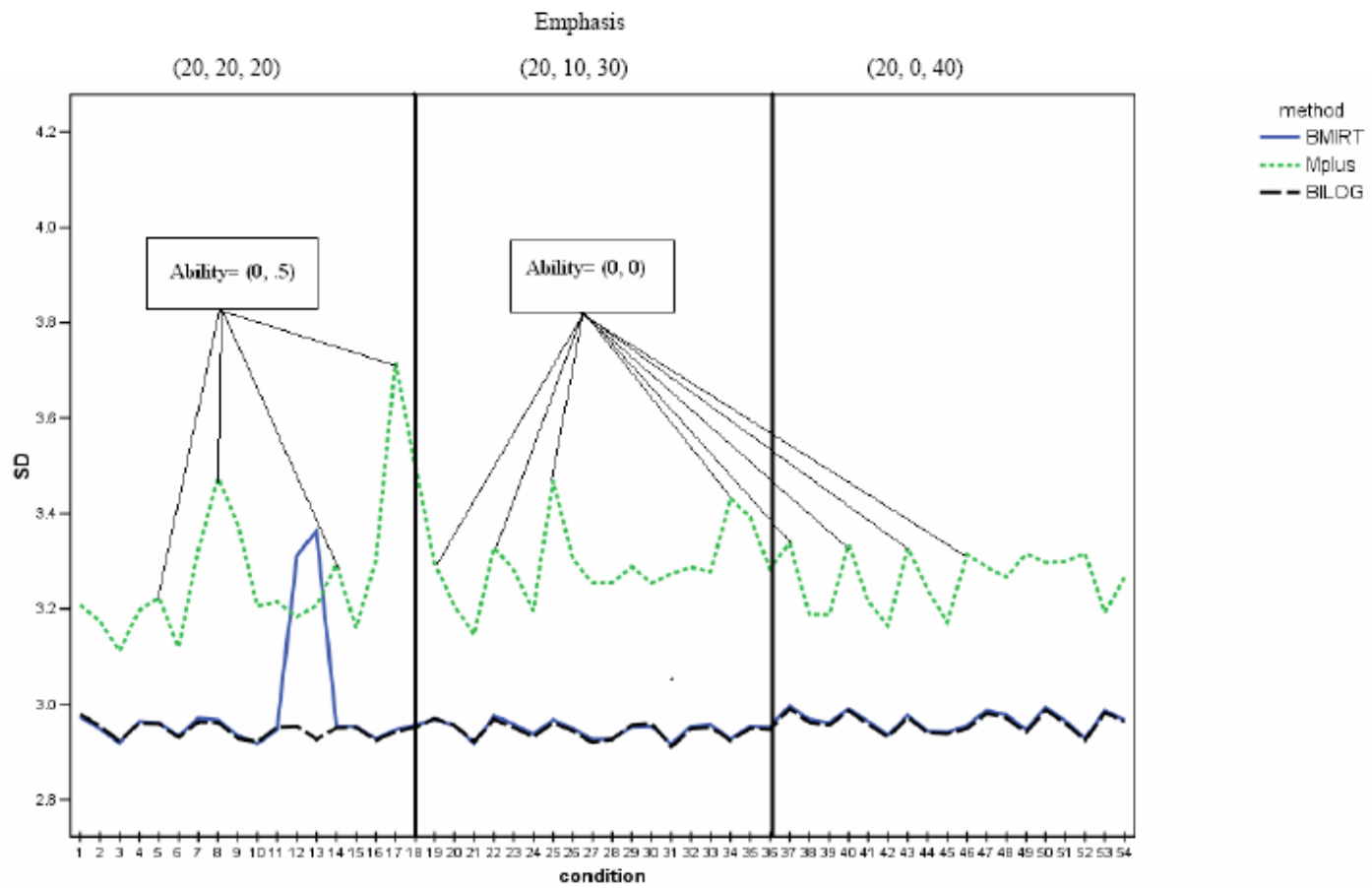


Figure 4-26 SD of the true score estimation in Group 2

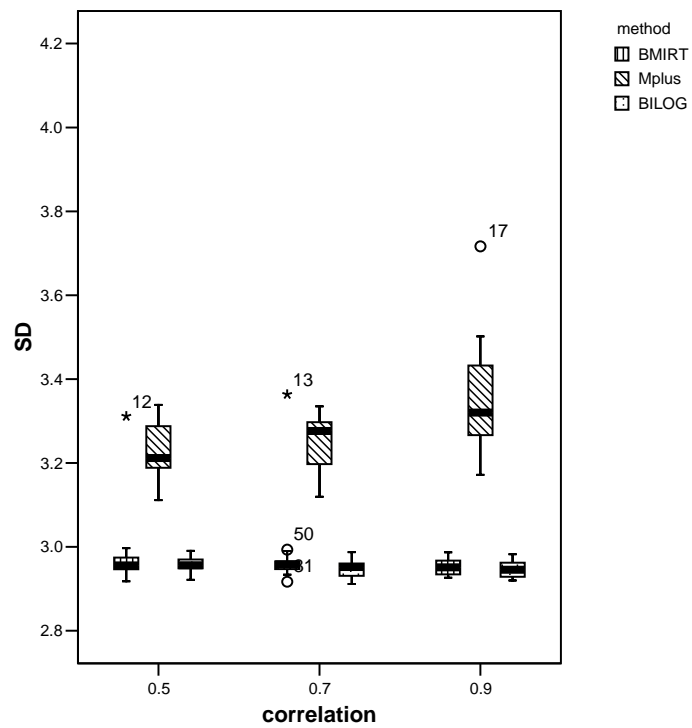
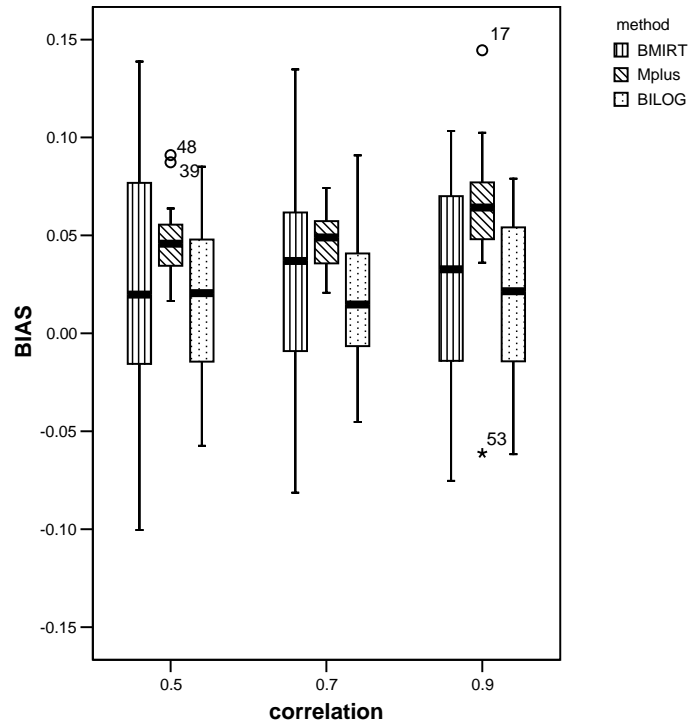


Figure 4-27 *BIAS* and *SD* of estimation of true score in Group 2 under three structural orthogonality levels

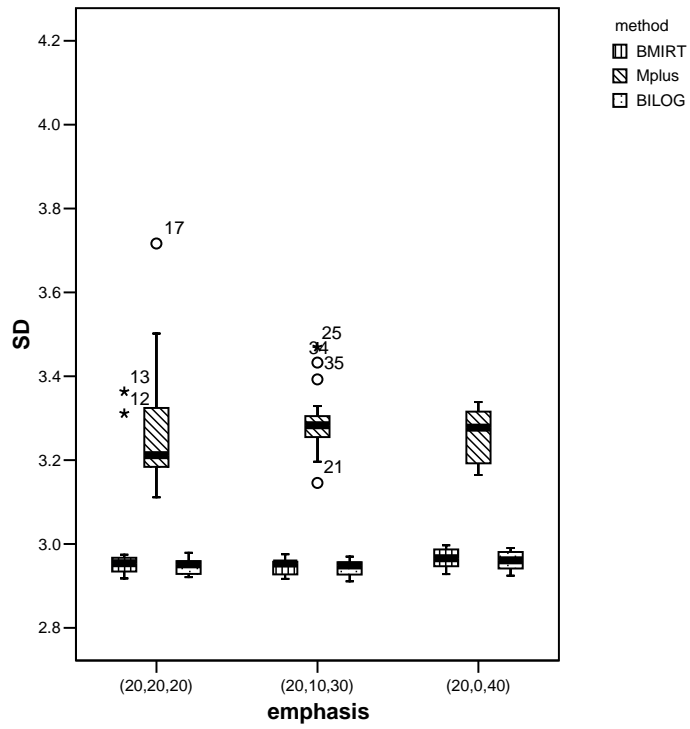
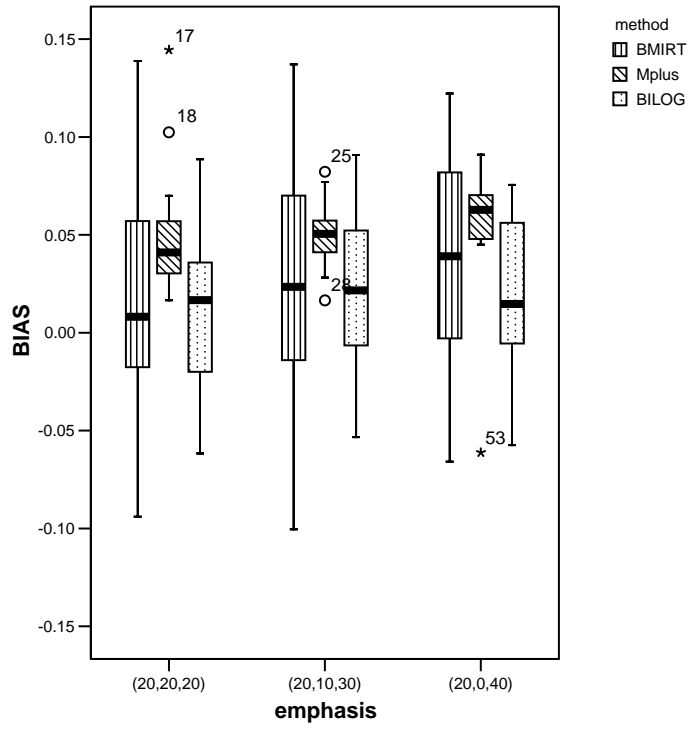


Figure 4-28 *BIAS* and *SD* of estimation of true score in Group 2 under three structural equivalence levels

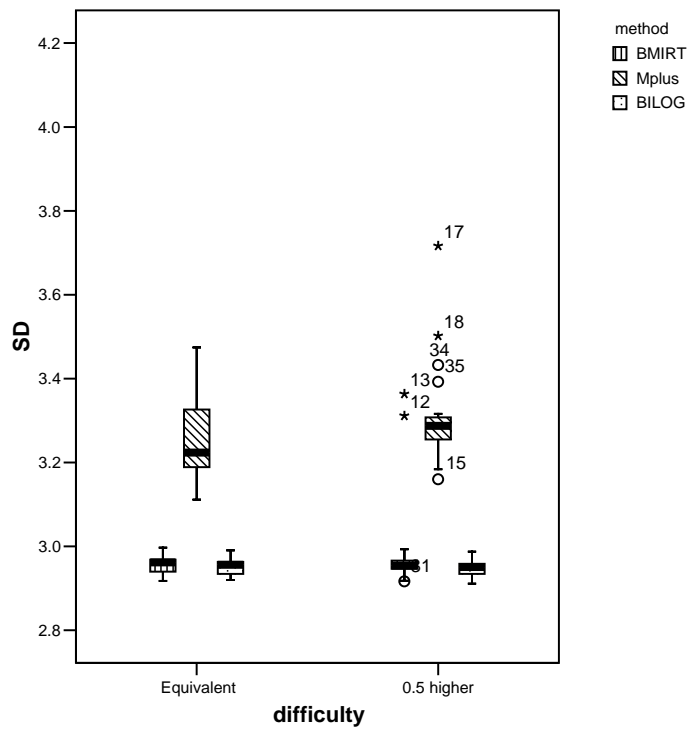
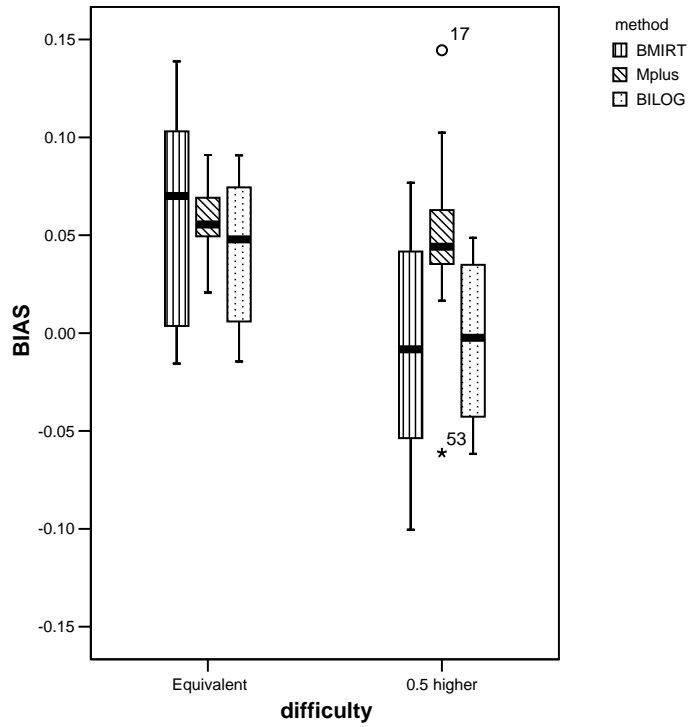


Figure 4-29 *BIAS* and *SD* of estimation of true score in Group 2 under two item difficulty equivalence levels

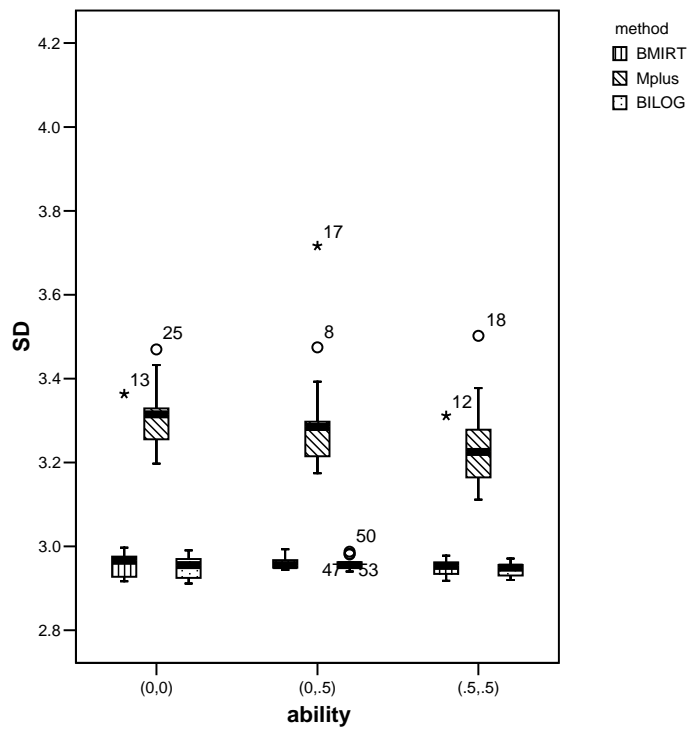
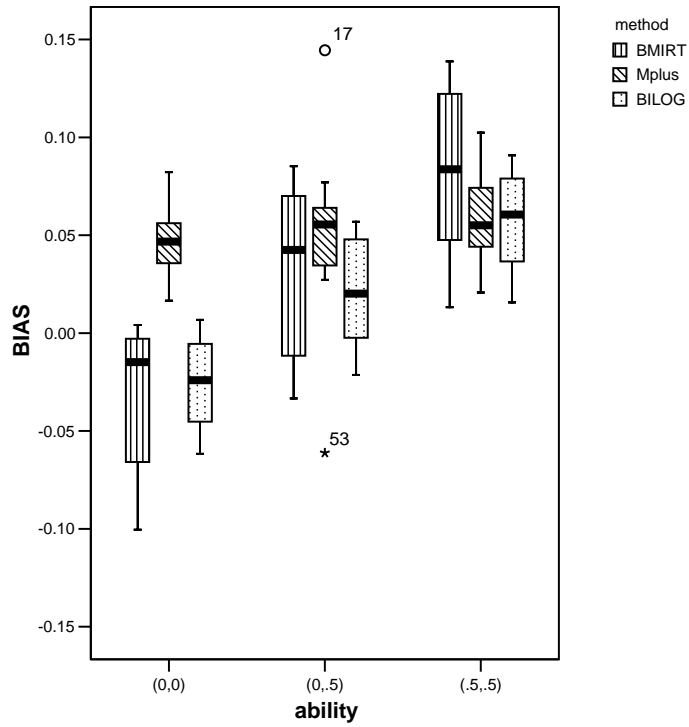


Figure 4-30 *BIAS* and *SD* of estimation of true score in Group 2 under three examinee group equivalence levels

## CHAPTER 5

### CONCLUSION AND DISCUSSION

#### 5.1 Summary of the Study

Multigroup analysis has been widely applied in educational measurement. In practice, most of the approaches for multigroup analysis are unidimensional. These approaches assume that the tests across groups measure a single uniform construct. This assumption, however, has been increasingly challenged. Not only the test construct within each group might be multidimensional, but the dimensions might change across groups. To solve this problem, multidimensional approaches have been proposed. Multidimensional IRT (MIRT) and factor analysis methods are two important ones. The purpose of this study was to investigate the performance of MIRT and factor analysis methods in analyzing multigroup multidimensional data. The performance of the unidimensional IRT method, compared with its multidimensional counterparts, was also investigated. The three multidimensional methods investigated were the concurrent MIRT calibration method, the separate MIRT calibration method with linking, and the concurrent factor analysis method. The unidimensional IRT method investigated was the concurrent unidimensional IRT calibration method.

The study was based on simulated data. A common item nonequivalent groups design was employed. There were two test forms. Each had 60 items, 20 of which were common items. Each form was developed to measure two abilities,  $\theta_1$  and  $\theta_2$ . The 20 common items measured both  $\theta_1$  and  $\theta_2$  and the 40 unique items measured only ability, either  $\theta_1$  or  $\theta_2$ . Assume that each test form was taken by a group of 2,000 imaginary examinees. Form 1 was taken by Group 1 and Form 2 was taken by Group 2. Four factors were manipulated to emulate real test conditions, including the

structural orthogonality, reflected by the correlation between  $\theta_1$  and  $\theta_2$ ; the equivalence of test structure, reflected by the number of items measuring  $\theta_1$  and  $\theta_2$  in each form; the equivalence of item difficulty; and the equivalence of examinee groups, reflected by the mean proficiency on  $\theta_1$  and  $\theta_2$  in each group. In total, there were 54 combinations of conditions. Under each condition 100 replications were made. The item response data was generated based on multidimensional compensatory two-parameter logistic (MC2PL) model. The concurrent MIRT calibration was carried out by the program BMIRT; the separate MIRT calibration was carried out by the program NOHARM, and the linking was done through a method modified from the method proposed by Min (2003); the concurrent factor analysis calibration was carried out by the program Mplus; the concurrent unidimensional IRT calibration was carried out by the program BILOG.

The performance of the calibration methods was evaluated based on the recovery of item parameters and the estimate of the true score of the examinees. The evaluation of the item parameter recovery was only available for the three multidimensional calibration methods. Two criteria were used: *BIAS* measured the mean difference between the estimated and true value of the parameters; *SD* reflected the stability of the estimation. The evaluation of the estimation of true score was only available for two of the three multidimensional calibration methods (the concurrent MIRT calibration method and the concurrent factor analysis calibration method) and the concurrent unidimensional IRT calibration method. Again, two criteria were used: *BIAS* measured the mean of the difference between the estimated true score and “true” true score of the examinees in each group; *SD* reflected the variability of the difference among the examinees in each group.



## 5.2 Summary of the Results

### 5.2.1 The Recovery of Item Parameters

From the results of the analysis, it can be concluded that the three multidimensional approaches performed differently with respect to the recovery of item parameters. The manipulated factors had some effect on the recovery of the parameters, but in a different way for each of the methods.

The key findings are as follows:

- (1) The bias of the estimate of  $a_1$  and  $d$  from the concurrent factor analysis method (Mplus) and the separate MIRT method (NOHARM) were comparable and very close to zero under most conditions. But the bias of the estimate of  $a_2$  from the two methods was less similar, with that from NOHARM fluctuating more widely across conditions.
- (2) For  $a_1$ ,  $a_2$ , and  $d$ , the bias of the estimate from the concurrent factor analysis method tended to be more consistent across conditions than the other two methods. The results indicated that the concurrent factor analysis method was less affected by the manipulated factors with respect to the bias of the estimation. However, the stability of the estimate from the concurrent factor analysis method fluctuated widely across conditions.
- (3) The concurrent MIRT calibration method tended to underestimate  $a_1$  and  $a_2$  and overestimate  $d$ . The estimate from the concurrent MIRT method usually had more bias than that of the other two methods under most conditions. However, it was more stable than the other two methods under most conditions and the stability was less affected by the manipulated factors.

- (4) When the correlation between the two dimensions increased, the estimate of  $a_1$  and  $a_2$  from the concurrent MIRT method became more biased, and that from all three methods became less stable. This effect was not evident in the estimate of  $d$ , except when the correlation was 0.9, then the estimate from the concurrent factor analysis method became less stable. A possible reason for this phenomenon might be that when the correlation between the dimensions was relatively high, it became more difficult for the analytic methods to define the difference between the dimensions and, therefore, it's harder to get unbiased and stable estimates of the parameters related to the dimensions. The parameter  $d$  reflects the overall difficulty of the item and is not directly related to the dimensions. Therefore, it was less affected by the correlation between dimensions.
- (5) For all three methods, the estimate of  $a_2$  had much more bias than that of  $a_1$  when the test structure of the two forms was not equivalent, specifically, when more emphasis was put on  $\theta_2$  in Form 2. This indicated that when the common items did not represent the dimensions equally well in a test, the estimate of the parameters related to the underrepresented dimension tended to be more biased. However, the stability of the estimate from all three methods was not negatively affected. The estimate of  $a_2$  from the concurrent factor analysis method became more stable when more emphasis was put on  $\theta_2$ . The following is a tentative explanation for this phenomenon. In the multigroup analysis, the common scale for the parameters from different tests is constructed based on the common items between the tests. When some dimension(s) can not be represented by the common items as well as the other dimensions, the common scale for the

underrepresented dimension(s) would be more biased and, therefore, the parameters on this common scale tend to be more biased. However, the bias of the common scale does not necessarily relate to the stability of the parameter estimation. In this study, when more items measured  $\theta_2$  in the test, more information about  $\theta_2$  was recovered from the data, and this, in turn, helped to get a more stable estimate of the parameters related to this dimension.

- (6) No effect of the equivalence of item difficulty between the two forms was found in the estimates of  $a_1$  and  $a_2$  from any of the three methods.

However, the estimate of  $d$  from the concurrent MIRT method tended to have more bias when the item difficulty of the two forms was not equivalent. However, this effect was not found in the estimates from the other two methods.

- (7) When the two examinee groups were not equivalent, the estimate of  $a_1$ ,  $a_2$ , and  $d$  from the separate MIRT method became more biased and less stable.

This effect was not found in the estimates from the other two methods.

Linking error might be a reason for such a change in the estimate from the separate MIRT method. As was discussed earlier, when the two groups were equivalent, no linking was made to the estimate of the parameters from different groups because they were already on the same scale. When the two groups were not equivalent, however, linking was needed. The increase in bias and decrease in stability of the parameter estimates when linking was conducted indicated that linking error might be a reason for such a change.

### 5.2.2 The Estimation of True Score of Examinees

With respect to the estimate of true score of examinees, the two multidimensional methods (the concurrent MIRT method and the concurrent factor analysis method) and the unidimensional IRT method performed differently with respect to bias and stability of the estimation. As to the recovery of the item parameters, the estimation of the true scores from different methods was also affected by the manipulated factors, but in somewhat different ways.

The followings are the key findings:

- (1) The estimate of true scores from all three methods had relatively small bias, compared with the range of total score.
- (2) In both groups, the estimate of true score from the concurrent MIRT method and the concurrent UIRT method were quite comparable, with respect to both bias and stability. In Group 1, the estimate of the true score from the two methods had very little bias. In Group 2, however, the estimate had much more bias and the bias tended to fluctuate a lot across conditions. One possible reason for the larger bias in Group 2 might be that during the process of concurrent estimation, Group 1 was treated as the reference group, where the joint distribution of the abilities is constrained to be standard multivariate normal and this happened to be true in Group 1. The distribution of the abilities in Group 2, however, was estimated based on the data. Therefore, the ability distribution in Group 1 had no estimation error, whereas that in Group 2 had error. As a result, the true score estimate in Group 2 had more bias than that in Group 1. The stability of the estimation in the two groups was comparable and consistent across conditions.

- (3) Compared with the two IRT methods, the estimate of true scores from the concurrent factor analysis method were less stable. In Group 1, the concurrent factor analysis method had more bias than the two IRT methods and the bias tended to fluctuate across conditions. In Group 2, however, the concurrent factor analysis method did not always have larger bias than the two IRT methods.
- (4) With respect to the effect of the correlation between the dimensions, no clear evidence regarding its impact was found in the estimates from the two IRT methods. The estimate from the concurrent factor analysis method became slightly more biased and less stable when the correlation between the dimensions increased. It was expected that the correlation between the dimensions would affect the estimates of the UIRT method in this way: the lower the correlation, the more bias and the less stable the estimates. However, such a pattern was not found in the results, which indicated that the UIRT method was robust to the multidimensionality of the data, at least under the conditions investigated in this study. In regard to the two multidimensional methods, although the correlation between the dimensions affected the estimation of the item parameters, the effect was much less in the estimation of the true score.
- (5) When the test structure of the two forms was not equivalent, the estimates of the true score in Group 1 from the concurrent factor analysis method tended to be more biased and less stable; in Group 2, the estimates from the concurrent MIRT and the concurrent factor analysis method tended to be slightly more biased. No evident effect was found in the estimate from the concurrent UIRT method, which indicated that the unidimensional

method was also robust to the nonequivalence of test structure.

- (6) When the two forms were equally difficult, all three methods tended to overestimate the true score in Group 2. However, when Form 2 was more difficult than Form 1, the two IRT methods no longer overestimated the true scores in Group 2. This change, however, was not found in the estimate from the factor analysis method.
- (7) When the two groups became less equivalent, specifically, when Group 2 had higher ability than Group 1 on one or both dimensions, the estimate of true scores in Group 1 from the two IRT methods was not affected appreciably, whereas that from the concurrent factor analysis method became slightly more biased and less stable under some conditions. In Group 2, however, the estimate from the two IRT methods became increasingly positively biased. Such a change was not evident in the estimate from the concurrent factor analysis method. This might indicate that the IRT methods somewhat “overreacted” to the increase in the ability of the examinees.

### 5.3 Discussion and Future Study

#### (1) Unidimensional vs. multidimensional methods

In estimating the true score of examinees, the performance of the concurrent unidimensional IRT method was quite comparable to its multidimensional counterpart with respect to both bias and stability of the estimates. The unidimensional IRT method was robust to the multidimensionality of the data under the conditions investigated in this study. It was also robust to the nonequivalence of the test structure across groups, when the test of different groups had different measurement emphases

on the same set of dimensions. This indicates that for a test, such as reading, where the dimensions or contents are moderately to highly related and the same set of dimensions or contents are measured across groups, applying the concurrent unidimensional IRT method in the multigroup analysis, such as equating or vertical scaling, might not be a problem. However, the results from this study are not readily generalized to the more complicated situation in which the correlation between dimensions is relatively low or the dimensions or contents change across groups. An example of this situation is the vertical scaling of a science test that spans a wide range of grades. More study is needed.

## (2) IRT vs. factor analysis methods

In regard to the recovery of the item parameters, the concurrent factor analysis method, in general, did a better job than the two MIRT methods. The bias of the estimate from the concurrent factor analysis method was comparable to, and sometimes smaller than, that from the separate MIRT method and it was always smaller than that from the concurrent MIRT method. The bias from the concurrent factor analysis method was more consistent across conditions and was less affected by the manipulated factors. Compared with the two MIRT methods, the estimates from the concurrent factor analysis method were also very stable, except when the correlation between dimensions were very high. This indicated that the concurrent factor analysis method might be a useful tool for item calibration in the multigroup analysis, which, in practice, is primarily done by the IRT methods. In general, one limitation of employing the factor analysis methods in item calibration is that they currently cannot model guessing (giving a correct response to an item by guessing) in the item response. When guessing is present, a preliminary analysis needs to be done

to adjust the response data before item calibration is conducted. The concurrent factor analysis method investigated in this study has this problem. However, guessing is not a problem for the two MIRT methods in this study because guessing is allowed in the model. How does the concurrent factor analysis method perform, compared with the MIRT methods, in item calibration when guessing is present? This question might be worth studying in the future.

With respect to the estimation of the true score, the two IRT methods, in general, performed better than the concurrent factor analysis method in Group 1, with smaller bias and more stability. Although the estimate of item parameters from the concurrent MIRT method had more bias than that from the concurrent factor analysis method, the underestimation of  $a_1$  and  $a_2$  and the overestimation of  $d$  seems to cancel each other out in the estimation of the true score. However, the two IRT methods tended to “overreact” to the change in the factors such as item difficulty and examinee ability. When the item difficulty in Form 2 increased, the estimate of the true score in Group 2 decreased more than it was supposed to. Similarly, when the examinees in Group 2 had higher ability than those in Group 1, the estimate of true score in Group 2 increased more than it was supposed to. The concurrent factor analysis method, however, was less affected by these factors. In summary, the results from this study indicated that the two IRT methods performed better than the factor analysis method in estimating the true score of the examinees. However, this conclusion can not be easily generalized to other conditions because the effect of the factors on the IRT methods needs to be studied further.



### (3) Concurrent vs. separate calibration methods

When the two groups were equivalent, the concurrent factor analysis method and the separate IRT method performed comparably, with respect to both estimation bias and stability. When the two groups were not equivalent, however, the parameter estimate from the separate MIRT method tended to be more biased and less stable than that from the concurrent factor analysis method. Linking error might contribute to this change. This might indicate that the concurrent factor analysis method is a better choice for the item calibration than the separate MIRT method when the examinee groups are not equivalent. On the other hand, because only one linking method was investigated in this study, one may ask how the other linking methods perform compared to this one under different conditions and what the most appropriate linking method is that could minimize the linking error. More investigation is needed to answer these questions.

### (4) The representation of the dimensions by the common items

When the test structure was not equivalent across groups, the selection of the common items would be a problem. If the common items can not represent the dimensions of a test equally well, the parameter(s) related to the underrepresented dimension(s) would be more biased. However, when the two tests do not have equivalent test structure, it's not possible for the common items to represent the dimensions equally well for all tests. Then one can ask how to select the common items so that they can minimize the overall bias in the parameter estimate resulting from the underrepresentation of dimensions. This too is worth further study.

## APPENDIX A

### Scale Transformation Method for separate MIRT calibration

In this study, the scale transformation for the separate MIRT calibration is performed as followed:

$$\mathbf{a}_i^{*'} = \mathbf{a}_i' \mathbf{E} \mathbf{K}$$

$$d_i^* = d_i + \mathbf{a}_i' \mathbf{E} \mathbf{m}$$

$$\boldsymbol{\theta}^* = \mathbf{K}^{-1}(\mathbf{E}\boldsymbol{\theta} - \mathbf{m})$$

where  $\boldsymbol{\theta}$  is  $n \times 1$  vector of ability parameters, where  $n$  is the number of dimensions;  $a_i$  and  $d_i$  are the estimate of item parameters on the compared scale;  $a_i^*$  and  $d_i^*$  are the estimate of the item parameters transformed from the compared scale to the base scale;  $\mathbf{K}$  is a  $n \times n$  diagonal dilation matrix;  $\mathbf{m}$  is a  $n \times 1$  translation vector for location;  $k$  is a central dilation constant for unit change; and  $\mathbf{E}$  is a  $n \times n$  identity matrix. Here, the matrix  $\mathbf{T}$  in Min (2003)'s method is replaced by the identity matrix  $\mathbf{E}$  so that the direction of the dimensions won't change in scale transformation.

$\mathbf{K}$  can be derived through the following procedure:

Assume  $\mathbf{A}_b$  is the item discrimination matrix for the common items in the base test, which is Form 1 in this study;  $\mathbf{A}_e$  is the item discrimination matrix for the common items in the equated test, which is Form 2 here.

$$\mathbf{A}_b = \mathbf{A}_e \mathbf{K} + \mathbf{E}$$

where  $\mathbf{E}$  is the residual matrix  $\mathbf{E} = \mathbf{A}_b - \mathbf{A}_e \mathbf{K}$ .  $\mathbf{K}$  can be derived by minimizing  $tr(\mathbf{E}'\mathbf{E})$ . In result,  $\mathbf{K} = \text{diag}[\mathbf{A}_b' \mathbf{A}_e \mathbf{T}] \times (\text{diag}(\mathbf{A}_e' \mathbf{A}_e))^{-1}$

$\mathbf{m}$  can be derived by the following procedure:

Assume  $\mathbf{D}_b$  is the item difficulty vector for the common items in the base test;  $\mathbf{D}_e$  is the item difficulty vector for the common items in the equated test.

$$\mathbf{D}_b = \mathbf{D}_e + \mathbf{A}_c \mathbf{m} + \mathbf{Q}$$

where  $\mathbf{Q}$  is the residual matrix  $\mathbf{Q} = \mathbf{D}_b - \mathbf{D}_e - \mathbf{A}_c \mathbf{m}$ .  $\mathbf{m}$  can be derived by minimizing  $tr(\mathbf{Q}'\mathbf{Q})$ . In result,  $\mathbf{m} = (\mathbf{A}'_c \mathbf{A}_c)^{-1} (\mathbf{A}'_c (\mathbf{D}_b - \mathbf{D}_e))$ .

APPENDIX B

Table B-1 *BIAS* of  $a_1$  estimated from the three methods under all 54 conditions

Correlation	Ability	Method	Emphasis					
			(20,20,20)		(20,10,30)		(20,0,40)	
			Difficulty					
			Equivalent	.5 higher	Equivalent	.5 higher	Equivalent	.5 higher
0.5	(0,0)	BMIRT	-0.0234	-0.0172	-0.0098	-0.0113	-0.0093	-0.0073
		Mplus	0.0064	0.0109	0.0140	0.0125	<b>0.0030</b>	<b>0.0054</b>
		NOHARM	<b>0.0052</b>	<b>0.0037</b>	<b>0.0059</b>	<b>0.0044</b>	-0.0150	-0.0137
	(0,.5)	BMIRT	-0.0237	-0.0238	-0.0173	-0.0113	-0.0114	-0.0084
		Mplus	0.0059	<b>0.0037</b>	0.0087	0.0123	0.0011	0.0028
		NOHARM	<b>0.0015</b>	0.0082	<b>-0.0007</b>	<b>0.0044</b>	<b>-0.0005</b>	<b>0.0011</b>
	(.5,.5)	BMIRT	-0.0209	-0.0236	-0.0079	-0.0106	-0.0115	-0.0114
		Mplus	0.0118	0.0155	0.0147	0.0152	<b>0.0002</b>	<b>0.0002</b>
		NOHARM	<b>0.0022</b>	<b>0.0049</b>	<b>0.0058</b>	<b>0.0049</b>	-0.0010	-0.0018
0.7	(0,0)	BMIRT	-0.0345	-0.0361	-0.0200	-0.0198	-0.0233	-0.0271
		Mplus	<b>0.0045</b>	0.0049	0.0128	0.0142	<b>0.0053</b>	0.0016
		NOHARM	0.0051	<b>0.0037</b>	<b>0.0057</b>	<b>0.0051</b>	-0.0238	<b>0.0013</b>
	(0,.5)	BMIRT	-0.0353	-0.0315	-0.0245	-0.0200	-0.0277	-0.0275
		Mplus	0.0073	<b>0.0049</b>	0.0101	0.0139	0.0013	0.0012
		NOHARM	<b>0.0019</b>	0.0053	<b>0.0015</b>	<b>0.0085</b>	<b>0.0009</b>	<b>0.0008</b>
	(.5,.5)	BMIRT	-0.0310	-0.0344	-0.0221	-0.0218	-0.0247	-0.0252
		Mplus	0.0107	0.0113	0.0119	0.0133	0.0033	0.0020
		NOHARM	<b>0.0019</b>	<b>0.0074</b>	<b>-0.0007</b>	<b>0.0025</b>	<b>0.0030</b>	<b>0.0017</b>
0.9	(0,0)	BMIRT	-0.0412	-0.0409	-0.0167	-0.0192	-0.0342	-0.0308
		Mplus	0.0166	0.0180	0.0125	0.0126	<b>0.0000</b>	<b>0.0017</b>
		NOHARM	<b>0.0098</b>	<b>0.0062</b>	<b>0.0090</b>	<b>0.0047</b>	-0.0368	-0.0339
	(0,.5)	BMIRT	-0.0366	-0.0358	-0.0227	-0.0181	-0.0328	-0.0350
		Mplus	0.0216	0.0216	<b>0.0097</b>	<b>0.0096</b>	<b>-0.0007</b>	-0.0032
		NOHARM	<b>0.0041</b>	<b>0.0068</b>	-0.0044	0.0101	0.0055	<b>0.0028</b>
	(.5,.5)	BMIRT	-0.0312	-0.0424	-0.0171	-0.0246	-0.0350	-0.0280
		Mplus	0.0127	0.0216	0.0156	0.0125	-0.0030	<b>0.0057</b>
		NOHARM	<b>-0.0125</b>	<b>0.0044</b>	<b>0.0071</b>	<b>-0.0115</b>	<b>0.0023</b>	0.0095

Note: The bold-faced numbers in the table indicate the methods that resulted in the smallest *BIAS* under each condition.

Table B-2 *SD* of  $a_1$  estimated from the three methods under all 54 conditions

correlationb	Ability	Method	Emphasis					
			(20,20,20)		(20,10,30)		(20,0,40)	
			Difficulty					
			Equivalent	.5 higher	Equivalent	.5 higher	Equivalent	.5 higher
0.5	(0,0)	BMIRT	<b>0.0715</b>	0.0751	0.0733	0.0728	<b>0.0699</b>	<b>0.0717</b>
		Mplus	0.0824	0.0852	<b>0.0700</b>	<b>0.0687</b>	0.0760	0.0784
		NOHARM	0.0734	<b>0.0750</b>	0.0761	0.0750	0.0735	0.0828
	(0,.5)	BMIRT	<b>0.0712</b>	<b>0.0725</b>	0.0749	0.0732	<b>0.0704</b>	<b>0.0714</b>
		Mplus	0.0826	0.0824	<b>0.0698</b>	<b>0.0683</b>	0.0787	0.0782
		NOHARM	0.0912	0.0903	0.0889	0.0885	0.0848	0.0844
	(.5,.5)	BMIRT	<b>0.0736</b>	<b>0.0738</b>	0.0734	0.0745	<b>0.0714</b>	<b>0.0714</b>
		Mplus	0.0870	0.0852	<b>0.0697</b>	<b>0.0715</b>	0.0781	0.0798
		NOHARM	0.0898	0.0971	0.0889	0.0902	0.0850	0.0865
0.7	(0,0)	BMIRT	<b>0.0766</b>	<b>0.0764</b>	0.0758	0.0791	<b>0.0749</b>	<b>0.0750</b>
		Mplus	0.0911	0.0908	<b>0.0722</b>	<b>0.0770</b>	0.0856	0.0877
		NOHARM	0.0798	0.0786	0.0803	0.0834	0.0817	0.0969
	(0,.5)	BMIRT	<b>0.0779</b>	<b>0.0781</b>	0.0760	0.0761	<b>0.0737</b>	<b>0.0747</b>
		Mplus	0.0926	0.0927	<b>0.0746</b>	<b>0.0758</b>	0.0851	0.0854
		NOHARM	0.1000	0.1018	0.1012	0.0995	0.0955	0.0952
	(.5,.5)	BMIRT	<b>0.0790</b>	<b>0.0758</b>	<b>0.0756</b>	0.0761	<b>0.0745</b>	<b>0.0753</b>
		Mplus	0.0932	0.0929	0.0757	<b>0.0752</b>	0.0863	0.0856
		NOHARM	0.1014	0.1004	0.0970	0.1003	0.0946	0.0961
0.9	(0,0)	BMIRT	<b>0.0855</b>	<b>0.0847</b>	<b>0.0910</b>	<b>0.0879</b>	<b>0.0837</b>	<b>0.0824</b>
		Mplus	0.1471	0.1487	0.1095	0.1135	0.1463	0.1477
		NOHARM	0.1094	0.1076	0.1150	0.1214	0.1896	0.1688
	(0,.5)	BMIRT	<b>0.0841</b>	<b>0.0840</b>	<b>0.0836</b>	<b>0.0852</b>	<b>0.0814</b>	<b>0.0872</b>
		Mplus	0.1392	0.1354	0.0978	0.1023	0.1256	0.1387
		NOHARM	0.1651	0.1660	0.1673	0.1719	0.1663	0.1683
	(.5,.5)	BMIRT	<b>0.0864</b>	<b>0.0841</b>	<b>0.0892</b>	<b>0.0868</b>	<b>0.0842</b>	<b>0.0806</b>
		Mplus	0.1526	0.1557	0.1125	0.1114	0.1478	0.1386
		NOHARM	0.1715	0.1690	0.1687	0.1705	0.1660	0.1644

Note: The bold-faced numbers in the table indicate the methods that resulted in the smallest *SD* under each condition.

Table B-3 *BIAS* for  $a_2$  estimated from the three methods under all 54 conditions

correlation	Ability	Method	Emphasis					
			(20,20,20)		(20,10,30)		(20,0,40)	
			Difficulty					
			Equivalent	.5 higher	Equivalent	.5 higher	Equivalent	.5 higher
0.5	(0,0)	BMIRT	-0.0198	-0.0204	-0.0354	-0.0363	-0.0697	-0.0679
		Mplus	0.0065	0.0012	-0.0323	-0.0359	-0.0342	-0.0337
		NOHARM	<b>0.0044</b>	<b>0.0021</b>	<b>0.0034</b>	<b>-0.0016</b>	<b>0.0248</b>	<b>0.0221</b>
	(0,.5)	BMIRT	-0.0219	-0.0207	-0.0245	-0.0352	-0.0659	-0.0728
		Mplus	0.0027	0.0042	-0.0285	-0.0344	<b>-0.0334</b>	<b>-0.0353</b>
		NOHARM	<b>-0.0015</b>	<b>-0.0025</b>	<b>0.0040</b>	<b>-0.0031</b>	-0.1011	-0.0995
	(.5,.5)	BMIRT	-0.0276	-0.0231	-0.0383	-0.0318	-0.0766	-0.0771
		Mplus	-0.0046	0.0047	-0.0348	-0.0321	<b>-0.0310</b>	<b>-0.0376</b>
		NOHARM	<b>-0.0033</b>	<b>-0.0017</b>	<b>-0.0073</b>	<b>0.0026</b>	-0.1012	-0.1000
0.7	(0,0)	BMIRT	-0.0407	-0.0419	-0.0623	-0.0570	-0.1094	-0.1020
		Mplus	0.0053	-0.0023	-0.0334	-0.0289	-0.0317	-0.0294
		NOHARM	<b>0.0025</b>	<b>0.0005</b>	<b>0.0040</b>	<b>0.0019</b>	<b>0.0290</b>	<b>0.0040</b>
	(0,.5)	BMIRT	-0.0451	-0.0488	-0.0581	-0.0603	-0.1039	-0.1089
		Mplus	<b>-0.0021</b>	<b>-0.0023</b>	-0.0328	-0.0292	<b>-0.0305</b>	<b>-0.0327</b>
		NOHARM	-0.0084	-0.0098	<b>-0.0094</b>	<b>-0.0114</b>	-0.1110	-0.1119
	(.5,.5)	BMIRT	-0.0421	-0.0416	-0.0578	-0.0610	<b>-0.1138</b>	<b>-0.1030</b>
		Mplus	<b>0.0026</b>	0.0049	-0.0306	-0.0342	-0.0348	-0.0279
		NOHARM	-0.0044	<b>-0.0015</b>	<b>-0.0031</b>	<b>-0.0013</b>	-0.1198	-0.1099
0.9	(0,0)	BMIRT	-0.0637	-0.0618	-0.1017	-0.0992	-0.1482	-0.1458
		Mplus	-0.0034	<b>0.0006</b>	-0.0351	-0.0348	<b>-0.0275</b>	-0.0323
		NOHARM	<b>-0.0018</b>	0.0006	<b>-0.0036</b>	<b>-0.0018</b>	0.0275	<b>0.0222</b>
	(0,.5)	BMIRT	-0.0700	-0.0728	-0.0954	-0.0955	-0.1357	-0.1443
		Mplus	<b>-0.0088</b>	<b>-0.0127</b>	<b>-0.0328</b>	<b>-0.0295</b>	<b>-0.0267</b>	<b>-0.0288</b>
		NOHARM	-0.0462	-0.0507	-0.0604	-0.0678	-0.1610	-0.1434
	(.5,.5)	BMIRT	-0.0733	-0.0691	-0.1043	-0.0988	-0.1437	-0.1523
		Mplus	<b>0.0022</b>	<b>-0.0006</b>	<b>-0.0330</b>	-0.0384	<b>-0.0256</b>	<b>-0.0285</b>
		NOHARM	-0.0304	-0.0471	-0.0613	<b>-0.0337</b>	-0.1535	-0.1469

Note: The bold-faced numbers in the table indicate the methods that resulted in the smallest *BIAS* under each condition.

Table B-4 *SD* for  $a_2$  estimated from the three methods under all 54 conditions

correlation	Ability	Method	Emphasis					
			(20,20,20)		(20,10,30)		(20,0,40)	
			Difficulty					
			Equivalent	.5 higher	Equivalent	.5 higher	Equivalent	.5 higher
0.5	(0,0)	BMIRT	<b>0.0731</b>	<b>0.0742</b>	0.0737	0.0740	0.0775	0.0783
		Mplus	0.0847	0.0848	<b>0.0617</b>	<b>0.0633</b>	<b>0.0637</b>	<b>0.0641</b>
		NOHARM	0.0761	0.0766	0.0734	0.0755	0.0719	0.0750
	(0,.5)	BMIRT	<b>0.0722</b>	<b>0.0713</b>	0.0748	0.0732	0.0760	0.0801
		Mplus	0.0841	0.0822	<b>0.0624</b>	<b>0.0617</b>	<b>0.0626</b>	<b>0.0629</b>
		NOHARM	0.0882	0.0885	0.0899	0.0894	0.0762	0.0793
	(.5,.5)	BMIRT	<b>0.0733</b>	<b>0.0724</b>	0.0718	0.0744	0.0785	0.0778
		Mplus	0.0852	0.0858	<b>0.0616</b>	<b>0.0626</b>	<b>0.0640</b>	<b>0.0619</b>
		NOHARM	0.0876	0.0934	0.0865	0.0896	0.0762	0.0767
0.7	(0,0)	BMIRT	<b>0.0759</b>	<b>0.0763</b>	0.0723	0.0762	0.0779	0.0792
		Mplus	0.0908	0.0922	<b>0.0656</b>	<b>0.0676</b>	<b>0.0670</b>	<b>0.0678</b>
		NOHARM	0.0794	0.0801	0.0773	0.0794	0.0741	0.0847
	(0,.5)	BMIRT	<b>0.0759</b>	<b>0.0741</b>	0.0773	0.0744	0.0806	0.0770
		Mplus	0.0910	0.0917	<b>0.0676</b>	<b>0.0671</b>	<b>0.0681</b>	<b>0.0658</b>
		NOHARM	0.1002	0.0968	0.1019	0.0988	0.0859	0.0825
	(.5,.5)	BMIRT	<b>0.0761</b>	<b>0.0762</b>	0.0760	0.0756	0.0795	0.0780
		Mplus	0.0901	0.0923	<b>0.0685</b>	<b>0.0661</b>	<b>0.0682</b>	<b>0.0676</b>
		NOHARM	0.0985	0.0987	0.1008	0.0975	0.0855	0.0846
0.9	(0,0)	BMIRT	<b>0.0846</b>	<b>0.0850</b>	<b>0.0858</b>	<b>0.0842</b>	<b>0.0818</b>	<b>0.0864</b>
		Mplus	0.1464	0.1421	0.0919	0.0939	0.0990	0.0994
		NOHARM	0.1084	0.1084	0.1028	0.1065	0.1212	0.1162
	(0,.5)	BMIRT	<b>0.0842</b>	<b>0.0812</b>	<b>0.0806</b>	<b>0.0786</b>	<b>0.0801</b>	<b>0.0853</b>
		Mplus	0.1301	0.1326	0.0856	0.0852	0.0877	0.0961
		NOHARM	0.1632	0.1601	0.1691	0.1596	0.1515	0.1428
	(.5,.5)	BMIRT	<b>0.0841</b>	<b>0.0827</b>	<b>0.0849</b>	<b>0.0871</b>	<b>0.0800</b>	<b>0.0836</b>
		Mplus	0.1503	0.1593	0.0945	0.0937	0.0996	0.0942
		NOHARM	0.1651	0.1693	0.1604	0.1760	0.1508	0.1424

Note: The bold-faced numbers in the table indicate the methods that resulted in the smallest *SD* under each condition.

Table B-5 *BIAS* for  $d$  estimated from the three methods under all 54 conditions

correlation	Ability	Method	Emphasis					
			(20,20,20)		(20,10,30)		(20,0,40)	
			Difficulty					
			Equivalent	.5 higher	Equivalent	.5 higher	Equivalent	.5 higher
0.5	(0,0)	BMIRT	0.0035	0.0221	0.0074	0.0350	0.0073	0.0558
		Mplus	<b>-0.0012</b>	<b>-0.0005</b>	0.0091	0.0129	0.0054	0.0059
		NOHARM	-0.0017	-0.0011	<b>0.0001</b>	<b>-0.0005</b>	<b>-0.0043</b>	<b>-0.0022</b>
	(0,.5)	BMIRT	0.0036	0.0252	0.0105	0.0453	0.0189	0.0753
		Mplus	-0.0041	<b>-0.0008</b>	0.0100	0.0198	<b>0.0018</b>	<b>0.0164</b>
		NOHARM	<b>-0.0008</b>	-0.0034	<b>0.0018</b>	<b>0.0100</b>	0.0033	0.0168
	(.5,.5)	BMIRT	0.0015	0.0220	0.0088	0.0364	0.0169	0.0700
		Mplus	-0.0100	-0.0087	<b>0.0035</b>	0.0083	0.0058	0.0065
		NOHARM	<b>-0.0015</b>	<b>-0.0033</b>	-0.0055	<b>-0.0035</b>	<b>-0.0008</b>	<b>0.0018</b>
0.7	(0,0)	BMIRT	0.0043	0.0215	0.0111	0.0350	0.0121	0.0679
		Mplus	-0.0018	-0.0012	0.0100	0.0124	0.0040	0.0121
		NOHARM	<b>0.0005</b>	<b>-0.0004</b>	<b>-0.0014</b>	<b>0.0014</b>	<b>-0.0017</b>	<b>0.0011</b>
	(0,.5)	BMIRT	<b>-0.0001</b>	0.0231	<b>0.0138</b>	0.0427	0.0199	0.0741
		Mplus	-0.0076	-0.0022	0.0148	0.0193	<b>0.0061</b>	<b>0.0120</b>
		NOHARM	-0.0051	<b>-0.0004</b>	0.0146	<b>0.0140</b>	0.0289	0.0309
	(.5,.5)	BMIRT	<b>0.0019</b>	0.0236	0.0121	0.0400	0.0257	0.0821
		Mplus	-0.0080	-0.0068	0.0063	0.0094	0.0060	0.0094
		NOHARM	-0.0039	<b>-0.0033</b>	<b>-0.0026</b>	<b>-0.0020</b>	<b>-0.0037</b>	<b>0.0004</b>
0.9	(0,0)	BMIRT	-0.0043	0.0187	0.0102	0.0426	0.0164	0.0972
		Mplus	-0.0102	-0.0049	0.0264	0.0128	0.0101	0.0362
		NOHARM	<b>-0.0032</b>	<b>-0.0026</b>	<b>-0.0022</b>	<b>-0.0017</b>	<b>-0.0019</b>	<b>-0.0036</b>
	(0,.5)	BMIRT	0.0003	0.0255	0.0147	0.0479	0.0251	0.1094
		Mplus	<b>-0.0001</b>	<b>0.0019</b>	<b>-0.0028</b>	<b>0.0155</b>	<b>-0.0022</b>	0.7327
		NOHARM	0.0003	0.0054	0.0378	0.0379	0.0826	<b>0.0861</b>
	(.5,.5)	BMIRT	0.0089	0.0215	0.0096	0.0444	0.0229	0.1187
		Mplus	<b>-0.0004</b>	-0.0085	0.0086	0.0153	<b>-0.0022</b>	0.0602
		NOHARM	0.0036	<b>-0.0034</b>	<b>-0.0051</b>	<b>-0.0026</b>	-0.0048	<b>-0.0005</b>

Note: The bold-faced numbers in the table indicate the methods that resulted in the smallest *BIAS* under each condition.



Table B-6 *SD* for *d* estimated from the three methods under all 54 conditions

correlation	Ability	Method	Emphasis					
			(20,20,20)		(20,10,30)		(20,0,40)	
			Difficulty					
			Equivalent	.5 higher	Equivalent	.5 higher	Equivalent	.5 higher
0.5	(0,0)	BMIRT	0.0694	0.0705	0.0718	0.0685	0.0701	0.0710
		Mplus	0.0785	0.0804	0.0756	0.0732	0.0747	0.0732
		NOHARM	<b>0.0613</b>	<b>0.0633</b>	<b>0.0629</b>	<b>0.0620</b>	<b>0.0613</b>	<b>0.0638</b>
	(0,.5)	BMIRT	<b>0.0711</b>	<b>0.0730</b>	<b>0.0714</b>	<b>0.0706</b>	<b>0.0706</b>	<b>0.0698</b>
		Mplus	0.0775	0.0796	0.0765	0.0768	0.0750	0.0763
		NOHARM	0.0766	0.0768	0.0789	0.0784	0.0740	0.0781
	(.5,.5)	BMIRT	<b>0.0725</b>	<b>0.0719</b>	<b>0.0714</b>	<b>0.0741</b>	<b>0.0691</b>	<b>0.0752</b>
		Mplus	0.0815	0.0814	0.0729	0.0762	0.0721	0.0788
		NOHARM	0.0762	0.0780	0.0768	0.0790	0.0759	0.0802
0.7	(0,0)	BMIRT	0.0707	0.0708	0.0747	0.0708	0.0698	0.0728
		Mplus	0.0743	0.0759	0.0765	0.0765	0.0771	0.0781
		NOHARM	<b>0.0612</b>	<b>0.0624</b>	<b>0.0632</b>	<b>0.0623</b>	<b>0.0604</b>	<b>0.0655</b>
	(0,.5)	BMIRT	<b>0.0747</b>	<b>0.0729</b>	<b>0.0724</b>	<b>0.0725</b>	<b>0.0725</b>	<b>0.0717</b>
		Mplus	0.0818	0.0814	0.0768	0.0750	0.0774	0.0815
		NOHARM	0.0761	0.0777	0.0754	0.0751	0.0759	0.0781
	(.5,.5)	BMIRT	<b>0.0757</b>	<b>0.0752</b>	<b>0.0755</b>	<b>0.0731</b>	<b>0.0728</b>	<b>0.0742</b>
		Mplus	0.0802	0.0827	0.0758	0.0761	0.0785	0.0792
		NOHARM	0.0812	0.0774	0.0759	0.0778	0.0770	0.0815
0.9	(0,0)	BMIRT	0.0754	0.0791	0.0766	0.0772	0.0747	0.0785
		Mplus	0.0936	0.0907	0.3784	0.0938	0.1016	0.1561
		NOHARM	<b>0.0610</b>	<b>0.0623</b>	<b>0.0605</b>	<b>0.0611</b>	<b>0.0623</b>	<b>0.0621</b>
	(0,.5)	BMIRT	0.0791	0.0772	0.0761	0.0762	0.0768	0.0845
		Mplus	0.0993	0.1048	0.1514	0.0994	0.1179	2.3621
		NOHARM	0.0741	0.0740	0.0758	0.0723	0.0724	<b>0.0763</b>
	(.5,.5)	BMIRT	0.0769	0.0802	0.0781	0.0788	0.0773	0.0802
		Mplus	0.0993	0.1068	0.1057	0.1518	0.1113	0.2326
		NOHARM	<b>0.0729</b>	<b>0.0722</b>	<b>0.0730</b>	<b>0.0751</b>	<b>0.0721</b>	<b>0.0765</b>

Note: The bold-faced numbers in the table indicate the methods that resulted in the smallest *SD* under each condition.

APPENDIX C

Table C-1 *BIAS* of true score estimated from the three methods

under all 54 conditions in Group 1

Correlation	Ability	Method	Emphasis					
			(20,20,20)		(20,10,30)		(20,0,40)	
			Difficulty					
			Equivalent	.5 higher	Equivalent	.5 higher	Equivalent	.5 higher
0.5	(0,0)	BMIRT	-0.0267	-0.0145	-0.0232	<b>-0.0009</b>	-0.0247	-0.0181
		Mplus	-0.0361	-0.0420	-0.1022	-0.0921	0.0371	0.0345
		BILOG	<b>-0.0156</b>	<b>-0.0129</b>	<b>-0.0122</b>	-0.0043	<b>-0.0121</b>	<b>-0.0175</b>
	(0,.5)	BMIRT	-0.0230	-0.0214	-0.0224	<b>0.0129</b>	-0.0347	-0.0031
		Mplus	-0.1130	-0.1010	-0.1027	-0.0728	0.0216	0.0468
		BILOG	<b>-0.0036</b>	<b>-0.0131</b>	<b>-0.0095</b>	0.0147	<b>-0.0211</b>	<b>-0.0012</b>
	(.5,.5)	BMIRT	-0.0137	<b>-0.0125</b>	-0.0120	<b>-0.0078</b>	-0.0119	-0.0043
		Mplus	-0.0979	-0.1067	-0.0843	-0.0943	0.0466	0.0466
		BILOG	<b>-0.0039</b>	-0.0142	<b>-0.0025</b>	-0.0087	<b>0.0008</b>	<b>-0.0013</b>
0.7	(0,0)	BMIRT	-0.0077	<b>-0.0047</b>	-0.0105	<b>-0.0048</b>	-0.0236	0.0071
		Mplus	-0.0125	-0.0286	-0.1165	-0.1186	0.0258	0.0438
		BILOG	<b>0.0008</b>	-0.0050	<b>-0.0028</b>	-0.0082	<b>-0.0147</b>	<b>0.0059</b>
	(0,.5)	BMIRT	-0.0173	-0.0107	-0.0157	0.0069	-0.0130	<b>0.0051</b>
		Mplus	-0.1100	-0.1309	-0.1160	-0.1079	0.0427	0.0452
		BILOG	<b>-0.0002</b>	<b>-0.0059</b>	<b>-0.0015</b>	<b>0.0085</b>	<b>-0.0004</b>	0.0076
	(.5,.5)	BMIRT	<b>0.0005</b>	<b>-0.0056</b>	-0.0209	<b>-0.0082</b>	-0.0128	<b>-0.0007</b>
		Mplus	-0.1111	-0.1284	-0.1237	-0.1202	0.0420	0.0377
		BILOG	0.0093	-0.0085	<b>-0.0134</b>	-0.0122	<b>-0.0023</b>	-0.0012
0.9	(0,0)	BMIRT	-0.0185	<b>-0.0128</b>	-0.0237	0.0031	-0.0240	<b>-0.0023</b>
		Mplus	<b>-0.0081</b>	-0.0164	-0.1236	-0.1160	0.0371	0.0521
		BILOG	-0.0117	-0.0181	<b>-0.0140</b>	<b>-0.0018</b>	<b>-0.0148</b>	-0.0026
	(0,.5)	BMIRT	-0.0109	<b>0.0009</b>	-0.0174	<b>0.0030</b>	-0.0278	-0.0086
		Mplus	-0.1245	-0.1084	-0.1293	-0.1117	0.0368	0.0108
		BILOG	<b>0.0026</b>	0.0033	<b>-0.0043</b>	0.0047	<b>-0.0148</b>	<b>-0.0056</b>
	(.5,.5)	BMIRT	-0.0093	<b>-0.0092</b>	-0.0146	<b>-0.0028</b>	-0.0188	<b>-0.0037</b>
		Mplus	-0.1183	-0.1072	-0.1028	-0.1127	0.0494	0.0404
		BILOG	<b>-0.0017</b>	-0.0123	<b>-0.0065</b>	-0.0111	<b>-0.0103</b>	-0.0106

Note: The bold-faced numbers in the table indicate the methods that resulted in the smallest *BIAS* under each condition.

Table C-2 *SD* of true score estimated from the three methods  
under all 54 conditions in Group 1

Correlation	Ability	Method	Emphasis					
			(20,20,20)		(20,10,30)		(20,0,40)	
			Difficulty					
			Equivalent	.5 higher	Equivalent	.5 higher	Equivalent	.5 higher
0.5	(0,0)	BMIRT	<b>2.871</b>	<b>2.864</b>	<b>2.866</b>	<b>2.866</b>	<b>2.862</b>	2.866
		Mplus	3.088	3.068	3.402	3.394	4.079	4.111
		BILOG	2.881	2.874	2.875	2.875	2.871	<b>2.875</b>
	(0,.5)	BMIRT	<b>2.865</b>	<b>2.872</b>	<b>2.864</b>	<b>2.863</b>	<b>2.881</b>	<b>2.856</b>
		Mplus	3.410	3.396	3.393	3.418	4.079	4.069
		BILOG	2.876	2.882	2.872	2.873	2.890	2.864
	(.5,.5)	BMIRT	<b>2.874</b>	3.198	<b>2.864</b>	<b>2.864</b>	<b>2.870</b>	<b>2.853</b>
		Mplus	3.399	3.430	3.415	3.408	4.102	4.073
		BILOG	2.885	<b>2.882</b>	2.874	2.873	2.878	2.864
0.7	(0,0)	BMIRT	<b>2.866</b>	3.200	<b>2.864</b>	<b>2.871</b>	<b>2.865</b>	<b>2.872</b>
		Mplus	3.119	3.079	3.370	3.391	3.969	3.974
		BILOG	2.871	<b>2.871</b>	2.870	2.876	2.872	2.878
	(0,.5)	BMIRT	<b>2.879</b>	<b>2.864</b>	<b>2.865</b>	<b>2.867</b>	<b>2.862</b>	<b>2.856</b>
		Mplus	3.396	3.378	3.406	3.379	3.975	3.943
		BILOG	2.885	2.871	2.869	2.873	2.866	2.862
	(.5,.5)	BMIRT	<b>2.865</b>	<b>2.854</b>	<b>2.863</b>	<b>2.865</b>	<b>2.875</b>	<b>2.870</b>
		Mplus	3.385	3.397	3.351	3.391	3.950	3.977
		BILOG	2.872	2.860	2.870	2.872	2.881	2.875
0.9	(0,0)	BMIRT	2.877	2.873	2.877	2.880	2.872	2.875
		Mplus	3.228	3.201	3.514	3.451	3.905	3.928
		BILOG	<b>2.873</b>	<b>2.868</b>	<b>2.873</b>	<b>2.877</b>	<b>2.869</b>	<b>2.873</b>
	(0,.5)	BMIRT	2.873	2.869	2.870	2.878	2.868	2.872
		Mplus	3.437	3.514	3.424	3.483	3.925	3.802
		BILOG	<b>2.869</b>	<b>2.867</b>	<b>2.867</b>	<b>2.876</b>	<b>2.866</b>	<b>2.868</b>
	(.5,.5)	BMIRT	2.874	2.877	2.871	2.879	2.871	2.881
		Mplus	3.481	3.497	3.494	3.473	3.955	3.895
		BILOG	<b>2.870</b>	<b>2.875</b>	<b>2.868</b>	<b>2.875</b>	<b>2.868</b>	<b>2.874</b>

Note: The bold-faced numbers in the table indicate the methods that resulted in the smallest *SD* under each condition.

Table C-3 *BIAS* of true score estimated from the three methods  
under all 54 conditions in Group 2

Correlation	Ability	Method	Emphasis					
			(20,20,20)		(20,10,30)		(20,0,40)	
			Difficulty					
			Equivalent	.5 higher	Equivalent	.5 higher	Equivalent	.5 higher
0.5	(0,0)	BMIRT	-0.0156	-0.0940	<b>0.0042</b>	-0.1004	<b>-0.0131</b>	-0.0659
		Mplus	0.0245	<b>0.0166</b>	0.0542	<b>0.0165</b>	0.0465	<b>0.0450</b>
		BILOG	<b>-0.0123</b>	-0.0435	0.0068	-0.0534	-0.0145	-0.0574
	(0,.5)	BMIRT	0.0665	-0.0334	0.0751	-0.0115	0.0819	0.0263
		Mplus	<b>0.0272</b>	0.0345	0.0552	0.0370	0.0638	0.0618
		BILOG	0.0343	<b>-0.0200</b>	<b>0.0523</b>	<b>0.0067</b>	<b>0.0479</b>	<b>0.0032</b>
	(.5,.5)	BMIRT	0.1388	<b>0.0133</b>	0.1371	0.0503	0.1222	0.0769
		Mplus	<b>0.0555</b>	0.0378	<b>0.0509</b>	0.0411	0.0910	0.0873
		BILOG	0.0851	0.0358	0.0799	<b>0.0366</b>	<b>0.0735</b>	<b>0.0364</b>
0.7	(0,0)	BMIRT	<b>0.0031</b>	<b>-0.0083</b>	-0.0090	-0.0814	<b>-0.0029</b>	-0.0531
		Mplus	0.0460	0.0357	0.0517	<b>0.0348</b>	0.0562	0.0479
		BILOG	0.0066	-0.0336	<b>-0.0065</b>	-0.0434	-0.0055	<b>-0.0452</b>
	(0,.5)	BMIRT	0.0547	-0.0316	0.0617	-0.0166	0.0852	0.0331
		Mplus	0.0509	0.0304	0.0478	0.0282	0.0703	0.0640
		BILOG	<b>0.0267</b>	<b>-0.0214</b>	<b>0.0391</b>	<b>-0.0037</b>	<b>0.0562</b>	<b>0.0137</b>
	(.5,.5)	BMIRT	0.1343	0.0407	0.1348	0.0531	0.1071	0.0450
		Mplus	<b>0.0207</b>	0.0442	<b>0.0573</b>	0.0501	0.0742	0.0688
		BILOG	0.0886	<b>0.0339</b>	0.0909	<b>0.0408</b>	<b>0.0724</b>	<b>0.0157</b>
0.9	(0,0)	BMIRT	<b>-0.0017</b>	-0.0753	<b>-0.0140</b>	-0.0624	0.0042	-0.0543
		Mplus	0.0570	<b>0.0361</b>	0.0822	0.0690	0.0692	<b>0.0469</b>
		BILOG	-0.0023	-0.0617	-0.0142	<b>-0.0421</b>	<b>0.0052</b>	-0.0490
	(0,.5)	BMIRT	0.0520	<b>-0.0176</b>	0.0700	<b>-0.0022</b>	0.0763	0.0228
		Mplus	0.0700	0.1445	0.0771	0.0595	<b>0.0558</b>	-0.0612
		BILOG	<b>0.0314</b>	-0.0199	<b>0.0542</b>	-0.0024	0.0569	<b>0.0116</b>
	(.5,.5)	BMIRT	0.1033	0.0570	0.1028	0.0427	0.0905	0.0476
		Mplus	<b>0.0691</b>	0.1024	<b>0.0480</b>	0.0436	<b>0.0547</b>	0.0795
		BILOG	0.0783	<b>0.0487</b>	0.0790	<b>0.0368</b>	0.0755	<b>0.0417</b>

Note: The bold-faced numbers in the table indicate the methods that resulted in the smallest *BIAS* under each condition.

Table C-4 *SD* of true score estimated from the three methods  
under all 54 conditions in Group 2

Correlation	Ability	Method	Emphasis					
			(20,20,20)		(20,10,30)		(20,0,40)	
			Difficulty					
			Equivalent	.5 higher	Equivalent	.5 higher	Equivalent	.5 higher
0.5	(0,0)	BMIRT	<b>2.974</b>	<b>2.918</b>	<b>2.969</b>	2.927	2.997	2.955
		Mplus	3.209	3.206	3.294	3.255	3.339	3.314
		BILOG	2.979	2.921	2.970	<b>2.927</b>	<b>2.991</b>	<b>2.949</b>
	(0,.5)	BMIRT	<b>2.949</b>	<b>2.947</b>	<b>2.956</b>	<b>2.953</b>	2.968	2.987
		Mplus	3.175	3.215	3.206	3.288	3.189	3.288
		BILOG	2.955	2.952	2.956	2.957	<b>2.962</b>	<b>2.981</b>
	(.5,.5)	BMIRT	<b>2.919</b>	3.311	<b>2.918</b>	<b>2.954</b>	2.962	2.978
		Mplus	3.111	3.184	3.146	3.254	3.188	3.267
		BILOG	2.922	<b>2.954</b>	2.921	2.960	<b>2.956</b>	<b>2.971</b>
0.7	(0,0)	BMIRT	2.964	3.364	2.976	2.917	2.990	2.947
		Mplus	3.197	3.207	3.329	3.274	3.335	3.316
		BILOG	<b>2.961</b>	<b>2.926</b>	<b>2.969</b>	<b>2.911</b>	<b>2.986</b>	<b>2.942</b>
	(0,.5)	BMIRT	2.962	2.954	2.959	2.954	2.964	2.993
		Mplus	3.224	3.294	3.282	3.287	3.215	3.297
		BILOG	<b>2.959</b>	<b>2.951</b>	<b>2.952</b>	<b>2.949</b>	<b>2.958</b>	<b>2.987</b>
	(.5,.5)	BMIRT	2.933	2.954	2.937	2.958	2.935	2.965
		Mplus	3.119	3.160	3.196	3.278	3.164	3.300
		BILOG	<b>2.930</b>	<b>2.952</b>	<b>2.931</b>	<b>2.952</b>	<b>2.931</b>	<b>2.960</b>
0.9	(0,0)	BMIRT	2.971	2.928	2.967	2.926	2.977	2.928
		Mplus	3.324	3.301	3.470	3.432	3.328	3.316
		BILOG	<b>2.963</b>	<b>2.925</b>	<b>2.962</b>	<b>2.924</b>	<b>2.973</b>	<b>2.924</b>
	(0,.5)	BMIRT	2.967	2.947	2.949	2.954	2.944	2.987
		Mplus	3.475	3.716	3.305	3.393	3.241	3.193
		BILOG	<b>2.962</b>	<b>2.942</b>	<b>2.944</b>	<b>2.950</b>	<b>2.940</b>	<b>2.982</b>
	(.5,.5)	BMIRT	2.934	2.957	2.927	2.953	2.942	2.967
		Mplus	3.377	3.502	3.255	3.284	3.172	3.266
		BILOG	<b>2.929</b>	<b>2.952</b>	<b>2.920</b>	<b>2.947</b>	<b>2.938</b>	<b>2.963</b>

Note: The bold-faced numbers in the table indicate the methods that resulted in the smallest *SD* under each condition.

## REFERENCES

- Ackerman, T. (1988, April). *An explanation of differential item functioning from a multidimensional perspective*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- Ackerman, T. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Educations*, 7(4), 255-278.
- Baker, F., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, 28, 147-162.
- Baker, F. (1992). *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker.
- Bartholomew, D. J. (1985, July). A unified view of factor analysis, latent structure analysis and scaling. Invited paper given at the 4-th European Meeting of the Psychometric Society and Classification Societies, Cambridge.
- Béguin, A., Hanson, B., & Glas, C. (2000). *Effect of multidimensionality on separate and concurrent estimation in IRT equating*. Paper presented at the American Educational Research Association, New Orleans, LA.
- Béguin, A., & Hanson, B. (2001). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.
- Bentler, P. (2004). *EQS Structural Equations Program*. Encino, CA: Multivariate Software, Inc.
- Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179-197.
- Bock, R., D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-445.
- Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, 12(3), 261-280.
- Bock, R. D., & Zimowski, M. (1996). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.
- Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (1999). *TESTFACT 3: Test scoring, items statistics, and full-information item factor analysis*. Chicago: Scientific Software International.

- Bollen, K.A.(1989). *Structural equations with latent variables*. New York: John Wiley.
- Bolt, M., & Lall, V. (2003). Estimation of compensatory and noncompensatory multidimensional Item Response models using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395-414.
- Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*, 5-32.
- Crocker, L. M. (1986). *Introduction to classical and modern test theory*. Orlando: Holt, Rinehart and Winston, Inc.
- Davey, T. (1991, June). *Some issues in linking multidimensional item calibrations*. Paper presented at the Office of Naval Research Contractors Meeting on Model-based Psychological Measurement, Princeton, NJ.
- Davey, T., Oshima, T., & Lee, K. (1996). *Linking multidimensional item calibrations*. *Applied Psychological Measurement, 20*, 405-416.
- Doody-Bogan, E., & Yen, W. (1983, April). *Detecting multidimensionality and examining its effect on vertical equating with the three parameter logistic model*. Paper presented at the annual meeting of the American educational Association, Montreal.
- Dickenson, T. S. (2005). Comparison of various ability estimates to the composite ability best measured by the total test score. Dissertation, University of South Carolina.
- Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 40*, 175-186.
- Finch, H. (2006). Comparison of the performance of varimax and promax rotations: factor structure recovery for dichotomous items. *Journal of Educational Measurement, 43*(1), 39-52.
- Fraser C., & McDonald R. P. (1988). *NOHARM: Least Squares Item Factor Analysis*. *Multivariate Behavioral Research, 23*, 267-269.
- Goldstain, H. (1983). Measuring changes in educational attainment over time: Problems and Possibilities. *Journal of Educational Measurement, 20*, 369-377.
- Glockner-Rist A., & Hoijtink, H (2003). The best of both worlds: factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling, 10*(4), 544-565.
- Haley, D.C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error (Technical Report 15)*. Stanford, CA: Stanford University, Applied Mathematics and Statistics laboratory.

- Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese psychological Research*, 22, 144-149.
- Hambleton, R.K., Swaminathan, H., & Roger, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publication.
- Hancock, G. R., Buehl, M., & Ployhart, R. E. (2002, April). Second-order latent growth models with shifting indicators. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Hanson, B., & Beguin, A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26(1), 3-24.
- Hanson, B., & Zeng, L. (2004). *ST: A Computer Program for IRT Scale Transformation*. Iowa Testing Program, the University of Iowa.
- Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2, 41-54.
- Horst, P. (1965). *Factor Analysis of Data Matrices*. New York: Holt, Rinehart and Winston.
- Hung, P., Wu, Y., & Chen, Y. (1991). *IRT item parameter linking: Relevant issues for the purpose of item banking*. Paper presented at the international Academic Symposium on Psychological Measurement, Tainan, Taiwan.
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36(4), 409-426.
- Jöreskog, K. G., & Sörbom, D. (2004). *LISREL 8.7 for Windows [Computer Software]*. Lincolnwood, IL : Scientific Software International, Inc.
- Kaskowitz, G., & De Ayala, R. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement*, 25(1), 39-52.
- Kim, S.-H., & Cohen, A. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement*, 29(1), 51-66.
- Kim, S.-H., & Cohen, A. (1998). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, 26(1), 25-41.
- Kim, S., & Kolen, M. (2003). *POLYST: A Computer Program for Polytomous IRT Scale Transformation*. Iowa Testing Program, the University of Iowa.



- Kim, S., & Kolen, M. (2004). *STUIRT: A Computer Program for Scale Transformation under Unidimensional Item Response Theory Models*. Iowa Testing Program, the University of Iowa.
- Kim, S. (2004). Unidimensional IRT scale linking procedures for mixed-format tests and their robustness to multidimensionality. Dissertation, University of Iowa, IA.
- Knol, D., & Berger, M. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, 26(3), 457-477.
- Kolen, M., & Brenna, R. (2004). *Test equating, Scaling, and Linking: Methods and Practices*. Springer Science+Business Media, Inc.
- Lee, K., & Oshima, T. (1996). IPLINK: Multidimensional and unidimensional item parameter linking in item response theory. *Applied Psychological Measurement*, 20, 230.
- Li, Y., & Lissitz, R. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement*, 24(2), 115-138.
- Lord, F., & Novick, M. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.
- Loyd, B., & Hoover, H. (1980). Vertical equating using the Rasch Model. *Journal of Educational Measurement*, 17, 179-193.
- Martineau, J. (2006). Distorting value added: the use of longitudinal, vertically scaled student achievement data for growth-based value-added accountability. *In press at Journal of Educational and Behavioral Statistics*.
- Martineau, J., etc. (2006). *Non-linear unidimensional scale trajectories through multidimensional content spaces: a critical examination of the common psychometric claims of unidimensionality, linearity, and interval-level measurement*. Paper presented at MARCES annual conference, College Park, MD
- Macro, G. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139-160.
- McCall, M. (2006). *Item response theory and longitudinal modeling: the real world is less complicated than we fear*. Paper presented to the MSDE/MARCES conference, College Park.
- McDonald, R. P. (1967). Nonlinear factor analysis (Psychometric Monographs, No. 15). Iowa City: Psychometric Society.
- McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 34, 100-117.

- McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, 6, 379-396.
- McDonald, R. P. (1985). *Factor Analysis and Related Methods*. Hillsdale, NJ: Erlbaum.
- Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43, 551-560.
- Muthén, B. (1979). A structural probit model with latent variables. *Journal of the American Statistical Association*, 24, 807-811.
- Muthén, B., Christofferson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, 46, 407-419.
- Muthén, B., Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: multiple-group and growth modeling in Mplus. Mplus Web Note. No. 4.
- Muthén, L., Muthén, B. (2006). *Mplus: Statistical analysis with latent variables*. Los Angeles, CA: Muthén & Muthén.
- Min, K.-S. (2003). *The Impact of Scale Dilation on the Quality of the Linking of Multidimensional Item Response Theory Calibrations*. Unpublished Dissertation, Michigan State University, East Lansing, MI.
- Mislevy, R. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, 11, 81-91.
- Mislevy, R., & Bock, R.D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models (2<sup>nd</sup> ed.)*. Mooresville, IN: Scientific Software.
- Muraki, E., & Bock, R. (1991). *PARSCALE: Parametric Scaling of Rating Data*. Chicago: Scientific Software International, Inc.
- Mulaik, S.A. (1972). *The foundations of factor analysis*. New York: McGraw Hill.
- Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review, Otaru University of Commerce*, 51(1), 1-23.
- Ogasawara, H. (2001a). Least squares estimation of item response theory linking coefficients. *Applied Psychological measurement*, 25(4), 3-24.
- Ogasawara, H. (2001b). Marginal maximum likelihood estimation of item response theory (IRT) equating coefficients for the common-examinee design. *Japanese Psychological Research*, 43(2), 72-82.

- Oshima, T., Davey, T., & Lee, K (2000). Multidimensional linking: four practical approaches. *Journal of Educational Measurement*, 37(4), 357-373.
- Reckase, M. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9, 401-412.
- Reckase, M., Ackerman, T., & Carlson, J. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, 25(3), 193-203.
- Reckase, M., & Mckinley, R. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, 15, 361-373.
- Reckase, M. (1995). A linear logistic multidimensional model for dichotomous items response data. In W. J. van der Linden and Hambleton (Ed.), *Handbook of Modern Item Response Theory* (pp. 271-286). New York: Springer.
- Reckase, M. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21(1), 25-36.
- Reckase, M., & Martineau, J. (2004). *The vertical scaling of science achievement tests*. Paper commissioned by the Committee on Test Design for K-12 Science Achievement, Center for Education, National Research Council.
- Roussos, L., Stout, W., & Marden, J. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, 35, 1-30.
- Rupp, A. & Zumbo, B.D. (2006). Understanding Parameter Invariance in Unidimensional IRT Models. *Educational and Psychological Measurement*, 66(1), 63-84.
- Skaggs, G. & Lissitz, R. W. (1988) Consistency of selected item bias indices: implications of another failure. American Educational Research Association
- Sorbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, 27, 229-239.
- Spence, P. (1996). *The effect of multidimensionality on unidimensional equating with item response theory*. Dissertation, University of Florida, FL.
- Stocking, M., & Lord, F. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
- Sympson, J. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive testing Conference* (pp. 82-98). Minneapolis: University of Minnesota.

- Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393-408.
- Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, 27(3), 159-203.
- Toit, M (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*, Scientific Software International, Inc.
- Tisak, J., & Meredith, W. (1990). Longitudinal factor analysis. *Statistical Methods in Longitudinal Research: Volume I, Principles and Structuring Change*, 125-149.
- Wang, M. (1986). *Fitting a unidimensional model to multidimensional item response data*. Paper presented at the Office of Naval Research contractors meeting.
- Way, W., & Tang, K. (1991). *A comparison of four logistic model equating methods*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Wingersky, M., Barton, M., & Lord, F. (1982). *LOGISTIC user guide*. Princeton, NJ: Educational Testing Service.
- Wingersky, M., & Lord, F. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, 8(3), 347-364.
- Yao, L. (2003). *BMIRT: Bayesian Multivariate Item Response Theory*. [Computer software]. Monterey, CA: CTB/McGraw-Hill.
- Yao, L., & Mao, X. (2004). *Unidimensional and multidimensional estimation of vertical scaled tests with complex structure*. Paper presented at the annual meeting of National Council on Measurement in Education, San Diego, CA.
- Yao, L. (2004). Bayesian Multivariate Item Response Theory and BMIRT software. *2004 Proceedings of the American Statistical Association, Statistical Computing Section [CD-ROM]*. Alexandria, VA: American Statistical Association.
- Yao, L. (2006). *A non-compensatory multidimensional multi-group IRT model for vertical scaling*. Paper presented at the National Council on Measurement in Education, San Francisco, CA.
- Yon, H., Reckase, M. (2005). *The impact of linking methodology on mirt parameter recovery in vertical scaling*. Paper presented at the International Meeting of Psychometric Society, Tilburg, NL

- Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software International, Inc.
- Zhang, J., & Stout, W. (1999). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, *64*, 129-152.
- Zimowski, M. (2003). Multiple-group analysis. *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*, 531-537. Chicago: Scientific Software Internal, Inc